

The functional independence measure 18-item version can be reported as a unidimensional interval-scaled metric: internal construct validity revisited

R. Maritz, A. Tennant, C. Fellinghauer, G. Stucki, Birgit Prodingler

Angaben zur Veröffentlichung / Publication details:

Maritz, R., A. Tennant, C. Fellinghauer, G. Stucki, and Birgit Prodingler. 2019. "The functional independence measure 18-item version can be reported as a unidimensional interval-scaled metric: internal construct validity revisited." *Journal of Rehabilitation Medicine* 51 (3): 193–200. <https://doi.org/10.2340/16501977-2525>.



THE FUNCTIONAL INDEPENDENCE MEASURE 18-ITEM VERSION CAN BE REPORTED AS A UNIDIMENSIONAL INTERVAL-SCALED METRIC: INTERNAL CONSTRUCT VALIDITY REVISITED

Roxanne MARITZ, MA^{1,2}, Alan TENNANT, PhD^{1,2}, Carolina FELLINGHAUER, PhD¹, Gerold STUCKI, MD^{1,2} and Birgit PRODINGER, PhD^{1,2,3}, on behalf of the NRP74 StARS clinics⁴

From the ¹Swiss Paraplegic Research, Nottwil, Switzerland, ²Department of Health Sciences and Health Policy, University of Lucerne, Lucerne, Switzerland, ³Faculty of Applied Health and Social Sciences, Technical University of Applied Sciences Rosenheim, Rosenheim, Germany and ⁴NRP74 StARS clinics (aarReha Schinznach – Stefan Bützberger, Berner Klinik Montana – Dieter Ernst, Berner Reha Zentrum Heiligenschwendi – Jürg Wägli, Clinica di Riabilitazione EOC Novaggio & Faido – Giovanni Rabito, Clinica Hildebrand – Sandro Foida, Clinique La Lignière – Nicolas Walther, Felix Platter Spital – Stefan Engelter, Hôpital du Valais Spital Wallis Centre Martigny, Sierre, Brig & Saint Amé – Els De Waele, Kantonsspital Baselland Bruderholz & Laufen – Beat Ritter, Klinik Schloss Mammern – Ruth Fleisch-Silvestri, Kliniken Valens Rehazentrum Valens, Rehazentrum Walenstadtberg & Rheinburg-Klinik – Stefan Bachmann, Rehaklinik Dussnang – Javier Blanco, Rehab Basel – Margret Hund – Georgiadis, Zürcher RehaZentrum Wald – Matthias Hermann, Rehaklinik Zihlschlacht – Michèle Bongetta, Spitäler Schaffhausen – Jan Kuchynka, Universitätsklinik Balgrist – Serge Altmann)

Objective: Since the 1990s the Functional Independence Measure (FIM™) was believed to measure 2 different constructs, represented by its motor and cognitive subscales. The practice of reporting FIM™ total scores, together with recent developments in the understanding of the influence of locally dependent items on fit to the Rasch model, raises the question of whether the FIM™ 18-item version can be reported as a unidimensional interval-scaled metric.

Design: Rasch analysis of the FIM™ using testlet approaches to accommodate local response dependency.

Patients: A calibration sample containing 946 cases of data from 11,103 patients undergoing neurological or musculoskeletal rehabilitation in Switzerland in 2016.

Results: Baseline analysis and the traditional testlet approach showed no fit with the Rasch model. When items were grouped into 2 testlets, fit to the Rasch model was achieved, indicating unidimensionality across all 18 items. A transformation table to convert FIM™ raw ordinal scores to the corresponding Rasch interval scaled values was created.

Conclusion: This study provides evidence that FIM™ total scores represent a unidimensional set of items, supporting their use in clinical practice and outcome reporting when applying the respective transformation table. This provides a basis for standardized reporting of functioning.

Key words: outcome assessment (healthcare); psychometrics; rehabilitation; activities of daily living; Rasch measurement model; Functional Independence Measure.

Accepted Jan 24, 2019; Epub ahead of print Mar 5, 2019

J Rehabil Med 2019; 51: 193–200

Correspondence address: Roxanne Maritz, Rehabilitation Services & Care Unit, Swiss Paraplegic Research, 6207 Nottwil, Switzerland. E-mail: roxanne.maritz@paraplegie.ch

The primary outcome of rehabilitation is functioning (1). In order to document and monitor functioning, existing data collection tools can be used (2).

LAY ABSTRACT

The aim of this study was to look in detail at the FIM™, an assessment tool often used for patients undergoing rehabilitation. Some users report the FIM™ as 2 scores: one related to motor tasks, the other to cognitive tasks; others recommend reporting it as a single score including both motor and cognitive tasks. This study explored whether it is statistically meaningful to sum all the points into a single FIM™ total score. The results support the current practice of summing the points into a single total score for patients undergoing musculoskeletal and neurological rehabilitation. The results also allowed an interval scale to be derived from the FIM™, enabling a broad range of calculations to be made using the FIM™ score, such as calculating the change in FIM™ outcomes from the time a patient is admitted to a rehabilitation clinic until their discharge.

The Functional Independence Measure (FIM™) is an assessment tool that is widely used in rehabilitation worldwide (3–6). The FIM™ is applied at the patient level to measure change throughout rehabilitation, at the institutional level to measure outcome quality, and at the national level for performance reporting or quality monitoring. Examples of use of the FIM™ are the reports of the Swiss National Association for Quality Development in Inpatient Care (ANQ) (7), the US model system for spinal cord injury (8) and traumatic brain injury (9), the Canadian Institute for Health Information (CIHI) (10), and the Australasian Rehabilitation Outcome Centre (AROC) (11). Furthermore, FIM™ measures can serve as a basis for inpatient rehabilitation payment (12). As with any assessment tool, in order to report valid total scores, certain psychometric standards must be met, including the assumption of unidimensionality. Furthermore, if an assessment tool is used to calculate change scores, it must be interval-scaled, rather than ordinal-scaled (6, 13). The Rasch measurement model can be used to examine assumptions such as unidimensionality or local item dependencies. Where satisfactory fit of data

to the model is achieved, an interval-scaled metric can be derived from ordinal scales (14, 15).

Earlier analysis of the FIM™ using Rasch analysis in the 1990s indicated that the FIM™ 18-item version incorporates 2 different constructs, represented by a motor scale and a cognitive scale, each of which should be scored separately (16). However, in clinical practice both the reporting of 2 separate motor and cognitive total scores and the reporting of a single total score of the FIM™, is evident (7, 9, 11). Since this first Rasch analysis of the FIM™, many others have been published, mostly on its motor subscale (17), but also on adaptations of the FIM™ (18, 19). More recently, the issue of so-called local item dependency has received attention (20). Local item dependency occurs when instrument items remain correlated when conditioned on the trait, what is functional independence in the case of the FIM™. Local dependency is indicated by significant correlation of the standardized analysis residuals. Fit of the FIM™ motor scale to the Rasch model has been shown to be seriously affected by local item dependency, which, once accommodated, resulted in adequate model fit (17).

Thus, given the recent methodological developments with regards to addressing the issue of local dependency in health scales, and inconsistency in reporting the FIM™ in practice, a review of the FIM™ 18-item version seemed appropriate, in order to address the following question: Is it possible to add all FIM™ items together to obtain a valid unidimensional total score, taking into account the local dependency in its item set? The objective of this study was therefore to revisit the question of whether the FIM™ can be reported as a unidimensional interval-scaled metric when local dependency is taken into account. Two specific aims in relation to the study's objective were: (i) to explore the metric properties of the FIM™; and (ii) to determine whether an interval-scale scoring system of the FIM™ 18-item version can be made available and, if so, to create an interval-scale transformation of the FIM™ raw scores when administered in the context of national quality monitoring in neurological and musculoskeletal rehabilitation.

METHODS

Subjects and setting

Data collected routinely for the Swiss national quality reporting, coordinated by the ANQ, was used for secondary analysis. All 64 Swiss rehabilitation clinics that provided data to the ANQ in 2016 for musculoskeletal or neurological rehabilitation were contacted, of which 30 voluntarily agreed to provide their ANQ datasets. Since the clinics can choose between different assessment tools in ANQ data collection, not all datasets contained FIM™ data. Thus, this study used datasets from 23 rehabilitation

clinics, with 11,103 complete cases in total, representative of 3 different Swiss language regions (German, French, Italian). The FIM™ was administered at admission and discharge. Ethics approval for the study was requested from the Swiss Ethics Commissions, which stated in a declaration of no objection that the project fulfils the general ethical and scientific standards for research with humans and poses no health hazards.

Functional Independence Measure

The FIM™ is an assessment tool comprising 18 items. Thirteen items belong to the motor subscale and 5 items belong to the cognitive subscale. All items are scored from 1 (total assistance) to 7 (complete independence). The FIM™ item scores are summed up to a total score, ranging between 18 and 126, or total motor score ranging between 13 and 91 and between 5 and 35 for the cognitive total score (4). The ANQ used German, French and Italian translations of the FIM™ based on its official English version, on which a translation agreement was made with the Uniform Data System for Medical Rehabilitation (UDSMR). As this is common practice, the translations have not been authenticated by the UDSMR. In order to qualify to administer the FIM™, the health professionals received training provided by the ANQ according to the respective UDSMR policy.

Sampling

A random stratified calibration sample was created using R (21), since type I errors, i.e. rejecting a hypothesis even if it was true, are likely to appear with a large sample size in Rasch analysis (22). The aim was to create a sample of approximately 1,000 cases, representing 4 equally sized subsamples, each with sufficient sample size for a stable item calibration and statistical interpretation (23, 24). Each subsample focused on one of the 2 different time-points of measurement, and one of the 2 different health condition groups of musculoskeletal and neurological rehabilitation: musculoskeletal cases at admission (MSKt1), musculoskeletal cases at discharge (MSKt2), neurological cases at admission (NEURt1) and neurological cases at discharge (NEURt2). To obtain precision across the whole range of scores (total score range 108; 18–126) and representation of language regions, a random sample was taken from each available total score per subsample and language region group. Cases that were selected from the admission subsamples were excluded and not selected for the discharge subsamples (25). Prior to the random selection all cases with missing values in a person's contextual factors of interest (described in more detail below) and all cases that scored an extreme score (18 or 126), were deleted, since they are excluded from the calculation of item difficulties by the Rasch measurement model. The sampling strategy is shown in Fig. 1.

Data analysis

To summarize basic sample characteristics and response distributions of the FIM™, descriptive statistics were conducted with Stata Version 14.2 (26). In order to achieve the study's first specific aim Rasch analysis was conducted using RUMM2030 (27). The analytical focus gave reference to local response dependency represented by residual correlations. High residual correlations indicate that items are measuring the same thing too closely (13). Furthermore, threshold disordering was examined, which indicates that the different response categories of an item are not in a successive order, i.e. do not represent an increasing level of functional independence. In addition, differential item

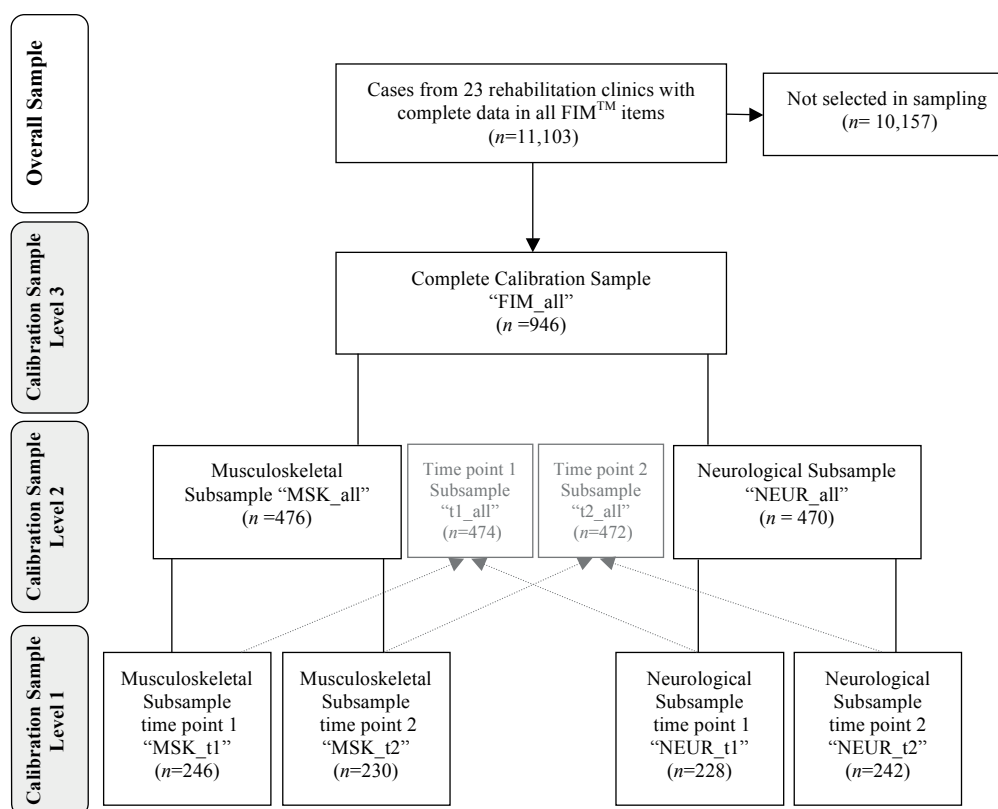


Fig. 1. Flow chart calibration sample with 3 different aggregation levels. FIM: Functional Independence Measure.

functioning (DIF) was evaluated, which indicates that, while accounting for the trait, an item works differently for certain groups defined by a contextual factor, such as gender or age. The partial credit model was applied, which has been shown previously to be the appropriate parametrization for the FIM™ (17, 28).

Baseline analyses

The baseline analysis tested how well the observed data from all 18 items fit the Rasch model (15). To do so, the individual and overall item-fit, the person-fit, the reliability indices α and person separation index (PSI), and the χ^2 p -value of the item-trait interaction standing for the fit of the data to the Rasch model were ascertained. The respective acceptable levels are represented in the bottom line of the corresponding results table. In addition, local response dependency among items was scrutinized, along with threshold disordering of item categories, and DIF for the following 7 factors: gender, age (4 age groups according to the interquartile ranges), nationality (Swiss or other), insurance (general, semi-private, private), rehabilitation group (neurological or musculoskeletal rehabilitation), clinic language (German, French or Italian) and time-point of measurement (admission t1, discharge t2). Both individual item-fit and DIF analyses p -values are Bonferroni adjusted in the RUMM2030 software.

Testlet approaches

Where the local independence assumption of the Rasch model was not met, testlet approaches were applied. A testlet is a simple sum score from a set of associated items, making the set into a

single new “super”-item in order to absorb their dependencies (20, 29–31). The creation of testlets revealed positive results in earlier Rasch analyses of the FIM™ motor scale (17). Two different testlet approaches were used: one approach, referred to as traditional testlet approach, creating testlets oriented at conceptually associated items and based on their residual correlations (32). By grouping similar items into super-items, such as, for example, all the transfer items of the FIM™, this traditional testlet approach highlights the potential differences, e.g. dimensionality between testlets unifies similar items, such as “self-care” or “transfer”. The other approach, referred to as alternative 2-testlet approach, divides conceptually similar items into 2 distinct testlets of equal size, taking alternative items in each testlet. This approach focuses on the total score of the FIM™ rather than the single items or groups of items by emphasizing the similarity of the items, as together they should measure the concept of functional independence. In delivering a bi-factor equivalent approach, the alternative 2-testlet approach has the advantage of creating testlets of equal size, as recommended by Andrich (29). Another advantage of the 2-testlet approach is that it allows for a conditional test of fit. Furthermore, all testlet-based approaches allow the calculation of the “explained common variance” attributable to the general “first factor”, indicating the proportion of variance retained to create a unidimensional latent estimate (29). Acceptable values of these additional statistics are indicated at the bottom of the respective testlet result table. The analysis of threshold disordering is not meaningful at the level of testlets, as a particular score can be derived in a number of ways, and is therefore not reported.

To ensure robustness of the results, the baseline analysis and the best-fitting testlet approach was conducted at 3 levels of aggregation of the calibration sample (see Fig. 1). In

Level 1 all 4 subsamples were analysed separately (MSKt1, MSKt2, NEURt1 and NEURt2). In Level 2 the rehabilitation group and time-point subsamples were aggregated respectively (MSKt1&t2, NEURt1&t2, t1MSK&NEUR, t2MSK&NEUR). Level 3 represents the aggregation of all 4 subsamples, i.e. the entire calibration sample (FIM_all). Together, these 3 aggregation levels resulted in 9 analysis steps.

For both testlet approaches, the emphasis is on making existing assessment tools work without the need to delete items or change the scoring structure.

Differential Item Functioning strategy

DIF was analysed in situations in which local dependencies could be accommodated satisfactorily with testlets. Where a lack of group invariance was observed, the testlets for the contextual factor were split on the basis of the strongest DIF, and continued until no further DIF was present (33). The split and unsplit solutions were then compared with each other on the basis of the Rasch person estimates, anchored to each other with an unsplit item free of DIF. An effect size calculation, based on the mean of the person estimates, their standard deviations, and the correlation of the split and unsplit version (34) was applied to determine whether DIF split was necessary for the final transformation table. If the effect size was below 0.2, DIF was considered small (35) and no action was taken to adjust for DIF.

Transformation table

The second specific aim of this study was to develop a transformation table in case fit to the Rasch model could be achieved. The solution with the best fit to the Rasch model was taken as a basis for this transformation, i.e. the solution with the most satisfactory core values for the entire calibration sample. The transformation table from FIMTM raw ordinal total scores to the corresponding interval-scaled values was based on the respective estimates according to the Rasch model.

RESULTS

Sample characteristics

The calibration sample included 946 cases. Of these, 476 were musculoskeletal cases and 470 neurological cases. A total of 474 cases were from time-point 1 admission, and 472 from time-point 2 discharge (see Fig. 1). FIMTM total scores had a mean of 81.7 (standard deviation (SD)=27.5, median=84). The mean age of subjects in the calibration sample was 71.6 years (SD=14.5, 20–102 years). The calibration sample was 43% (n=403) male and 57% (n=543) female; 41% (n=392) were from the German-speaking region of Switzerland, 25% (n=238) from the French-speaking region and 34% (n=316) from the Italian-speaking region; 84% (n=798) of the sample were Swiss and 16% (n=148) had another nationality. Insurance status was: 67% (n=633) general, 18% (n=172) semi-private, and 15% (n=141) private.

Baseline Rasch analysis

In the 9 baseline analysis steps across the 3 aggregation levels of the calibration sample, no fit to the Rasch model was achieved (Table I). In all analyses the p-values of the item-trait χ^2 were significant. Furthermore, in all analysis steps there were items that showed local dependencies among each other, DIF and threshold disordering. Information on threshold disordering and local dependency of the baseline analyses are shown in Appendix S1¹.

¹<https://doi.org/10.2340/16501977-2525>

Table I. Functional Independence Measure (FIMTM) baseline analyses

Sample	n/CI	Item-fit residuals Mean (SD)	Person-fit residuals Mean (SD)	χ^2 p-value	PSI	α	DIF (items)	Paired t-test (Lower ci %), %
MSK_t1	246/4	0.193 (2.496)	-0.183 (1.304)	0.000	0.961	0.967	age (M), language (A, B, D, F, L, R)	9.8 (0.0)
MSK_t2	230/4	0.098 (2.191)	-0.165 (1.359)	0.000	0.966	0.968	language (B, D, F, L, N, P)	17.4 (0.0)
MSK_all	476/8	0.193 (3.255)	-0.155 (1.280)	0.000	0.963	0.967	gender (Q), age (L, N), language (B, C, D, F, H, L, M, N, Q, R), time-point (L, M, N, O)	16.2 (0.0)
NEUR_t1	228/4	-0.046 (3.559)	-0.314 (1.745)	0.000	0.964	0.972	language (Q)	17.1 (14.3)
NEUR_t2	242/4	-0.461 (3.449)	-0.358 (1.595)	0.000	0.964	0.973	No DIF	15.3 (12.5)
NEUR_all	470/8	-0.369 (4.919)	-0.349 (1.678)	0.000	0.963	0.972	language (D, F, M, N, P, Q), time-point (L)	15.3 (13.3)
t1_all	474/8	0.101 (4.274)	-0.239 (1.609)	0.000	0.96	0.968	age (F, I, J, N, Q, R), language (B, D, F, L, N, Q), rehab-group (C, E, K, M, O, P, Q, R)	12.9 (10.9)
t2_all	472/8	-0.284 (3.957)	-0.293 (1.553)	0.000	0.964	0.971	language (B, D, M, N, Q), rehab-group (C, E, K, L, O, P, Q)	13.1 (11.2)
FIM_all	946/10	-0.077 (5.779)	-0.265 (1.609)	0.000	0.962	0.969	gender (L), age (N, O), language (B, D, F, H, L, M, N, Q, R), nationality (Q), insurance (O), time-point (L, M), rehab-group (C, E, K, L, M, O, P, Q, R)	11.1 (9.7)
Acceptable values		SD < 1.4	SD < 1.4	> 0.01	> 0.7	> 0.7	No DIF present	At least Lower ci < 5

MSK: musculoskeletal rehabilitation; NEUR: neurological rehabilitation; t1: admission; t2: discharge; all: combination of time-points and/or 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.

Testlet approaches

Within the traditional testlet approach 3 different versions of testlet combinations were applied, based on the underlying subscale structure of the FIM™. Two versions included 4 testlets for the motor scale, structured according to the FIM™ subtopics (self-care, sphincter control, transfers, locomotion) together with 2 combinations of the cognitive items. In one version all the cognitive FIM™ items were unified in one testlet, since they all showed local dependency among each other at the baseline analysis, resulting in a total of 5 testlets. In the other version, the cognitive items were split thematically according to the FIM™ subtopics into 2 testlets, communication and social cognition, resulting in a total of 6 testlets. The third version attempted to form similar sized testlets and was oriented at the residual correlations between the items and formerly reported clusters of the FIM™ (29, 36). In this version, 3 testlets were created: a self-care testlet incorporating items A–H, a mobility testlet incorporating items I–L, and a cognitive testlet incorporating items M–R. None of the 3 traditional testlet approaches, the 3-testlet, the 5-testlet and the 6-testlet version, resulted in fit to the Rasch model (see Table II).

In contrast, the alternative 2-testlet approach (with Testlet1 containing items A, C, E, G, I, K, M, O and Q, and Testlet2 containing items B, D, F, H, J, L, N, P and R) showed fit to the Rasch model across all 9 analyses steps. The *p*-values from the item-trait χ^2 were all non-significant at the 0.01 level, the reliability indexes all above 0.9, and the item- and person-fit estimates

within the set acceptable values. The expected common variance values retained in the latent estimate were all just above 1, indicating some marginal remaining residual local dependency among the testlets. The fit of all testlet solutions is summarized in Table II, and the application of the 2-testlet approach to all aggregation levels of the calibration sample is shown in Appendix S2¹.

Differential Item Functioning strategy

Despite overall fit, some DIF remained in the 2-testlet solution for the whole calibration sample. For eliminating all DIF, the successful 2-testlet solution of the whole calibration sample had to be split twice. Testlet2 first had to be split by rehabilitation group. Secondly, the group of musculoskeletal rehabilitation from Testlet2 had to be split into the 2 time-points, i.e. admission and discharge. This resulted in the following super-items: Testlet1, Testlet2_NEUR, Testlet2_MSkt1, and Testlet2_MSkt2. Testlet1 was the anchor for the comparison of the person estimates of the split and the unsplit version. The effect size calculation resulted in 0.11 (see Appendix S3¹), indicating that there was no need to split the final interval-scale transformation into different subgroups.

Transformation table

Based on the 2-testlet solution, an interval-based transformation table was created for all available FIM™ total scores, which can be used to transfer the

Table II. Testlet solutions on the level of the whole calibration sample (FIM_all)

<i>n</i> /CI	Testlets (items)	Item-fit residuals Mean (SD)	Person-fit residuals Mean (SD)	χ^2 <i>p</i> -value	PSI	α	DIF (Testlet)	A	Paired <i>t</i> -test, %	Cond. test of fit CI based
946/10	6 Testlets: Self-Care (A-F), Sphincter Control (G-H), Transfers (I-K), Locomotion (L-M), Communication (N-O), Social Cognition (P-R)	-0.156 (5.077)	-0.426 (1.200)	0.000	0.906	0.887	gender (T6), age (T2), language (T1, T2, T3, T4, T5, T6), insurance (T1, T5), time-point (T2, T4, T5), rehab-group (T1, T3, T4, T5, T6)	0.942	1.27	Only available for the 2-testlet approach
946/10	5 Testlets: Self-Care (A-F), Sphincter Control (G-H), Transfers (I-K), Locomotion (L-M), Cognition (N-R)	-0.010 (7.046)	-0.360 (1.138)	0.000	0.895	0.878	age (T2, T5), language (T1, T2, T3, T4, T5), nationality (T5), insurance (T1), time-point (T2, T4), rehab-group (T3, T4, T5)	0.930	1.16	
946/10	3 Testlets: Self-Care (A-H), Mobility (I-M), Cognition (N-R)	-1.419 (6.894)	-0.502 (1.049)	0.000	0.838	0.859	gender (T2, T3), age (T1), language (T1, T2, T3), nationality (T1, T3), insurance (T1), time-point (T2), rehab-group (T2, T3)	0.871	1.27	
946/10	2-testlets: Testlet1 (A, C, E, G, I, K, M, O, Q), Testlet2 (B, D, F, H, J, L, N, P, R)	-0.208 (0.317)	-0.614 (1.003)	0.408	0.980	0.981	rehab-group (T1, T2)	1.019	4.97	0.607
Acceptable values		SD <1.4	SD <1.4	>0.01	>0.7	>0.7	No DIF	>0.9	<5.00	>0.01

FIM all: Functional Independence Measure; all: combination of time-points and rehabilitation-groups; *n*: sample size; CI: class intervals; SD: standard deviation; PSI: person separation index; α : Cronbach's alpha; A: explained common variance; DIF: differential item functioning.

Table III. Functional Independence Measure (FIM™) total score transformation table: original scores to interval scores

Original FIM™ score	Rasch estimate	Transformed interval score	Original FIM™ score	Rasch estimate	Transformed interval score
18	-6.279	18.0	73	0.169	75.7
19	-5.686	23.3	74	0.232	76.3
20	-5.281	26.9	75	0.295	76.9
21	-5.005	29.4	76	0.358	77.4
22	-4.784	31.4	77	0.422	78.0
23	-4.594	33.1	78	0.485	78.6
24	-4.423	34.6	79	0.548	79.1
25	-4.263	36.0	80	0.612	79.7
26	-4.112	37.4	81	0.677	80.3
27	-3.966	38.7	82	0.741	80.9
28	-3.824	40.0	83	0.806	81.4
29	-3.686	41.2	84	0.871	82.0
30	-3.552	42.4	85	0.937	82.6
31	-3.420	43.6	86	1.004	83.2
32	-3.291	44.8	87	1.071	83.8
33	-3.166	45.9	88	1.139	84.4
34	-3.043	47.0	89	1.208	85.0
35	-2.923	48.0	90	1.277	85.6
36	-2.807	49.1	91	1.347	86.3
37	-2.693	50.1	92	1.418	86.9
38	-2.582	51.1	93	1.491	87.6
39	-2.473	52.1	94	1.564	88.2
40	-2.367	53.0	95	1.638	88.9
41	-2.264	53.9	96	1.714	89.6
42	-2.163	54.9	97	1.791	90.3
43	-2.064	55.7	98	1.869	90.9
44	-1.968	56.6	99	1.949	91.7
45	-1.874	57.4	100	2.030	92.4
46	-1.783	58.3	101	2.113	93.1
47	-1.693	59.1	102	2.197	93.9
48	-1.605	59.8	103	2.283	94.7
49	-1.519	60.6	104	2.371	95.4
50	-1.435	61.4	105	2.461	96.2
51	-1.353	62.1	106	2.553	97.1
52	-1.273	62.8	107	2.647	97.9
53	-1.194	63.5	108	2.742	98.8
54	-1.116	64.2	109	2.840	99.6
55	-1.041	64.9	110	2.941	100.5
56	-0.966	65.6	111	3.043	101.5
57	-0.893	66.2	112	3.148	102.4
58	-0.821	66.9	113	3.256	103.4
59	-0.750	67.5	114	3.365	104.3
60	-0.680	68.1	115	3.478	105.4
61	-0.611	68.7	116	3.593	106.4
62	-0.543	69.4	117	3.712	107.4
63	-0.476	70.0	118	3.834	108.5
64	-0.409	70.6	119	3.963	109.7
65	-0.343	71.1	120	4.098	110.9
66	-0.278	71.7	121	4.245	112.2
67	-0.213	72.3	122	4.410	113.7
68	-0.149	72.9	123	4.605	115.4
69	-0.085	73.5	124	4.853	117.7
70	-0.021	74.0	125	5.225	121.0
71	0.043	74.6	126	5.784	126.0
72	0.106	75.2			

ordinal-scaled FIM™ raw scores into interval-scaled FIM™ scores (see Table III).

DISCUSSION

This is the first study to provide evidence of the unidimensionality of the FIM™ 18-item version when administered to neurological and musculoskeletal patients in an inpatient rehabilitation setting. Although the

baseline analyses and the traditional testlet approaches did not result in Rasch model fit, an alternative 2-testlet approach, emphasizing the sameness of the FIM™ items with 2 equally sized super-items, achieved model fit. The robustness of these results was confirmed in repeating the same strategy for all subsets of the calibration sample. These results provide evidence for the internal construct validity of the FIM™ total scores based on 18 items, and thus support its reporting as a total score in clinical practice. Based on the current results, an interval-scale transformation table of the FIM™ total scores for use in national quality monitoring for neurological and musculoskeletal patients could be provided.

The traditional testlet approach applied in this study builds on the successful Rasch analysis strategy for FIM™ motor items (17). Nevertheless, when adding the cognitive item set to the motor items, the 4 testlet solution in the original paper could not be confirmed in this study. Highlighting the sameness of all the items through the 2-testlet approach attained model fit. The assumption of multidimensionality is often pursued in FIM™-related Rasch analyses (36). The extent of local dependency among the 18 items, clustered into the underlying structures raises questions as to whether the FIM™ should *a priori* be divided into multidimensional concepts, as proposed by Linacre and colleagues (16). The successful summation of the 18 FIM™ items in the current study places emphasis on the higher order construct of functional independence, incorporating both motor and cognitive aspects. Likewise, the Rasch analysis performed supports the theory that, when activities of daily living are observed, motor activities reflect some cognitive aspects and vice versa (18, 37). From a clinical perspective, the FIM™ offers different levels of granularity for reporting. Scores can be reported at the level of the single items, the item headings, e.g. sphincter control, on the level of the motor and cognitive subscales, or the level of the overall 18-item summary of functional independence. Thus, different granular levels of reporting are available, depending upon the use required.

The study can be also be seen as initial evidence that the German, French and Italian translations of the FIM™ do not substantially differ from each other, given the absence of substantial DIF by language. Furthermore, this study provides first evidence for the internal construct validity of the FIM™ 18-item version for musculoskeletal patients, given that no substantial DIF was present between the musculoskeletal and the neurological rehabilitation group. Nevertheless, the use of cognitive items for a musculoskeletal patient population within a national outcome report, as in the ANQ, remains debatable, and care should be

taken to consider whether only motor items should be assessed. In the US model system for burn injury, for example, only the FIM™ motor subscale is assessed (38). However, with regards to the ageing population and related comorbidity (39), both subscales may be of interest in musculoskeletal patients.

This study has the limitations of secondary data analysis. For example, there is a lack of information on the accuracy and consistency of the data-entry process, the selection of DIF factors was limited to the variables of the dataset and the use of non-validated translations. However, this design enabled a well-tailored calibration sample to be obtained from a large sample size. Another limitation is in the 2-testlet approach, which provides the basis for the transformation table. On the one hand, this approach was successful in attaining model fit. On the other hand, the approach does not allow a statement to be made about the hierarchy and difficulty of single items or a conceptually related group of items, since it focuses on the whole construct being measured through the assessment tool. However, while, for the purpose of quality or outcome reports, the FIM™ is based on the total score or change scores, data collection is still conducted on an item level, which allows clinicians to gain insight into the development of a single patient in a certain item or group of items, or to conduct a quality check of scores at the item level if the FIM™ was, for example, applied within a payment system.

The analysis of threshold disordering is also not possible with the testlet approach. Indeed, there is some evidence that disordered thresholds can themselves be caused by local dependency (20, 40). For example, if items are analysed within their subscales, threshold ordering may appear correct, but become disordered when subscales are summated together. Thus, it becomes impossible to determine if disordered thresholds are a consequence of local dependency, as the solution for local dependency renders interpretation of traditional thresholds invalid (20). Since the 2-testlet approach is a relatively new one for health assessment tools, further investigations are needed to confirm the influence of local dependency on such matters. However, this approach has the advantage that the total scores of a well-established and widely used assessment tool, such as the FIM™, can be converted on an interval-scale level, without deleting or re-scoring items.

We recommend the use of the interval transformation table provided in this study for neurological and musculoskeletal patients for national rehabilitation quality monitoring, in order to be able to calculate interval-scaled patient change scores for the FIM™, compared with its original ordinal scoring system (6). If the total scores are available in a digital format, as in the ANQ datasets, transformation can be implemented

easily in an electronic information system, by simply re-coding the total scores according to the table provided in the results. This interval scoring system has the advantage that it provides an important basis for the application of a standardized reporting system for functioning information (2, 41) in which the FIM™ could be integrated as a widely used instrument in rehabilitation. This is beneficial, as the standardized reporting of functioning information enables clinicians to continue using currently implemented assessment tools while also being able to compare and aggregate the information within and across tools, institutions or even countries. One caveat to this is that the interval-scale transformation is actually measured with error, as can be seen in its logit form in Appendix S4¹.

In conclusion, the results of this study support the internal construct validity of the FIM™ 18-item version and, consequently, the reporting of its total score, by applying the interval-scaled transformation table provided in this study. The fact that all the variance could be accommodated in the final estimate suggests that previous reports of multidimensionality may have been driven by a breach of the local independence assumption. This supports the intention of its developers and the way the FIM™ scores are used in clinical practice and in institutional and national monitoring. It is recommended to use the interval-scale transformation of the FIM™ total score for national quality monitoring for neurological and musculoskeletal patients, in order to adequately report change scores in patients' functioning. Furthermore, interval transformation provides a basis for integrating the FIM™ into a standardized reporting system for functioning information.

ACKNOWLEDGEMENTS

The authors thank Dr L. Menzi, Head of Rehabilitation ANQ, and K. Schmitt, Corporate Development Director of the Swiss Paraplegic Centre, for their good advice, provision of valuable information and 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 NRP74 StARS Project "Enhancing continuous quality improvement and supported clinical decision making by standardized reporting of functioning."

The authors have no conflicts of interests to declare.

REFERENCES

1. Stucki G, Bickenbach J. Functioning: the third health indicator in the health system and the key indicator for rehabilitation. *Eur J Phys Rehabil Med* 2017; 53: 134–138.
2. Stucki G, Prodinger B, Bickenbach J. Four steps to follow when documenting functioning with the International Classification of Functioning, Disability and Health. *Eur J Phys Rehabil Med* 2017; 53: 144–149.

3. Haigh R, Tennant A, Biering-Sorensen F, Grimby G, Marincek C, Phillips S, et al. The use of outcome measures in physical medicine and rehabilitation within Europe. *J Rehabil Med* 2001; 33: 273–278.
4. Keith RA, Granger CV, Hamilton BB, Sherwin FS. The Functional Independence Measure: a new tool for rehabilitation. *Adv Clin Rehabil* 1987; 1: 6–18.
5. Fiedler RC, Granger CV. The Functional Independence Measure: a measurement of disability and medical rehabilitation. In: Chino N. MJL, editor. *Functional evaluation of stroke patients*. Tokyo: Springer-Verlag; 1996, p. 75–92.
6. Uniform Data System for Medical Rehabilitation. The FIM(R) instrument: its background structure and usefulness. Buffalo: UDSMR; 2012 [cited 2018 Jul 19]. Available from: https://www.udsmr.org/Documents/The_FIM_Instrument_Background_Structure_and_Usefulness.pdf.
7. ANQ Nationaler Verein für Qualitätsentwicklung in Spitälern und Kliniken. Review Information: Rehabilitation. 2018 [cited 2018 Jul 19]. Available from: <https://www.anq.ch/en/departments/rehabilitation/>.
8. Model System Knowledge Translation Center. The Spinal Cord Injury Model System (SCIMS)-SCIMS PowerPoint presentation. 2013 [cited 2018 May 25]. Available from: <http://www.msktc.org/About-Model-Systems#SCI>.
9. Model System Knowledge Translation Center. The Traumatic Brain Injury Model Systems-TBIMS PowerPoint Presentation. 2017 [cited 2018 May 25]. Available from: <http://www.msktc.org/tbi>.
10. Canadian Institute for Health Information. National Rehabilitation Reporting System-Privacy Impact Assessment. 2015 [cited 2018 Nov 05]. Available from: https://www.cihi.ca/en/nrs_pia_en.pdf.
11. Australasian Rehabilitation Outcomes Centre. The FIM™ instrument in Australia. [cited 2018 Nov 05]. Available from: <https://ahsri.uow.edu.au/aroc/whatisfim/index.html>.
12. Carter GM, Relies DA, Ridgeway GK, Rimes CM. Measuring functioning for Medicare inpatient rehabilitation payment. *Health Care Financ Rev* 2003; 24: 25–44.
13. Andrich D. Rating scales and Rasch measurement. *Expert Rev Pharmacoecon Outcomes Res* 2011; 11: 571–585.
14. Ehlan A, Kucukdeveci A, Tennant A. The Rasch measurement model. In: F. F, editor. *Research issues in physical & rehabilitation medicine*. Pava: Maugeri Foundation; 2010, p. 89–102.
15. 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 Rheum* 2007; 57: 1358–1362.
16. Linacre JM, Heinemann AW, Wright BD, Granger CV, Hamilton BB. The structure and stability of the Functional Independence Measure. *Arch Phys Med Rehabil* 1994; 75: 127–132.
17. Lundgren Nilsson A, Tennant A. Past and present issues in Rasch analysis: the Functional Independence Measure (FIM) revisited. *J Rehabil Med* 2011; 43: 884–891.
18. Cantagallo A, Carli S, Simone A, Tesio L. MINDFIM: a measure of disability in high-functioning traumatic brain injury outpatients. *Brain Inj* 2006; 20: 913–925.
19. Medvedev ON, Turner-Stokes L, Ashford S, Siegert RJ. Rasch analysis of the UK Functional Assessment Measure in patients with complex disability after stroke. *J Rehabil Med* 2018; 50: 420–428.
20. Wilson M. Detecting and interpreting local item dependence using a family of Rasch models *Appl Psychol Meas* 1988; 12: 353–364.
21. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2018 [cited 2018 Nov 05]. Available from: <https://www.R-project.org/>.
22. Smith AB, Rush R, Fallowfield LJ, Velikova G, Sharpe M. Rasch fit statistics and sample size considerations for polytomous data. *BMC Med Res Methodol* 2008; 8: 33.
23. Linacre JM. Sample size and item calibration (or person measure) stability. *Arch Rasch Meas SIG*; 1994 [cited 2018 Nov 05]. Available from: <https://www.rasch.org/rmt/rmt74m.htm>.
24. Hagell P, Westergren A. Sample size and statistical conclusions from test of fit to the Rasch model according to the Rasch Unidimensional Measurement Model (RUMM) program in health outcome measurement. *J Appl Meas* 2016; 17: 416–431.
25. Mallinson T. Rasch analysis of repeated measures. *Rasch Meas Trans* 2011; 251: 1317.
26. StataCorp. Stata Statistical Software: Release 14. College Station, Texas: StataCor LP; 2015 [cited 2018 Nov 05]. Available from: <http://www.stata.com>.
27. Andrich D, Sheridan B, Luo G. RUMM2030 Version 5.4 for Windows. Rumm Laboratory Pty Ltd; 2012 [cited 2018 Nov 05]. Available from: <http://www.rummlab.com.au/>.
28. Masters GN. A Rasch model for partial credit scoring. *Psychometrika* 1982; 47: 149–174.
29. Andrich D. Components of variance of scales with a bifactor subscale structure from two calculations of alpha. *Educ Meas* 2016; 35: 25–30.
30. Wainer H, Kiely G. Item clusters and computer adaptive testing: a case for testlets. *J Educ Meas* 1987; 24: 185–202.
31. Tuerlinckx F, De Boeck P. The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychol Methods* 2001; 6: 181–195.
32. Christensen KB, Makransky G, Horton M. Critical values for Yen's Q3: identification of local dependence in the Rasch model using residual correlations. *Appl Psychol Meas* 2017; 41: 178–194.
33. Andrich D, Hagquist C. Real and artificial differential item functioning in polytomous items. *Educ Psychol Meas* 2015; 75: 185–207.
34. Morris SB, DeShon RP. Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychol Methods* 2002; 7: 105–125.
35. Cohen J. *Statistical power analysis for the behavioral sciences*. Hillsdale: Lawrence Erlbaum Associates; 1988.
36. Pretz CR, Kean J, Heinemann AW, Kozlowski AJ, Bode RK, Gebhardt E. A multidimensional Rasch analysis of the Functional Independence Measure based on the National Institute on Disability, Independent Living, and Rehabilitation Research Traumatic Brain Injury Model Systems National Database. *J Neurotrauma* 2016; 33: 1358–1362.
37. Tesio L, Valsecchi MR, Sala M, Guzzon P, Battaglia MA. Level of activity in profound/severe mental retardation (LAPMER): a Rasch-derived scale of disability. *J Appl Meas* 2002; 3: 50–84.
38. Model System Knowledge Translation Center. Burn injury resources. 2018 [cited 2018 Jul 16]. Available from: <https://msktc.org/burn/burn-resources>.
39. World Health Organization. World report on ageing and health Geneva, Switzerland: 2015.
40. Hammond A, Tennant A, Tyson SF, Nordenskiöld U, Hawkins R, Prior Y. The reliability and validity of the English version of the Evaluation of Daily Activity Questionnaire for people with rheumatoid arthritis. *Rheumatology* 2015; 54: 1605–1615.
41. Proding B, Tennant A, Stucki G. Standardized reporting of functioning information on ICF-based common metrics. *Eur J Phys Rehabil Med* 2018; 54: 110–117.