

Appendix S3. Differential Item Functioning Strategy for 2-testlet approach on the level of the whole calibration sample (FIM_all).

Analysis Name (Testlets)	DIF		Uniform (U) Non-uniform (N)	p-values	Interpretation
	Person Factor	Testlet			
TwoTestlet (T1, T2)	reha-group	T1	Uniform & Non-uniform	0.000037 0.000236	artificial, since it disappears in Split1
	reha-group	T2	Uniform & Non-uniform	0.000015 0.000416	lowest p-value, basis for following split
Split 1 (T1, T2_MSK, T2_NEUR)	age	T2_NEUR	Uniform	0.005376	T1 could serve as anchor
	language	T2_MSK	Non-uniform	0.000302	smallest but similar p-value than time-point in the same item
	time-point	T2_MSK	Uniform	0.000381	in order to reduce number of items (3 languages), time-point T2_MSK was used as basis for Split2
Split2 (T1, T2_NEUR, T2_MSK_t1, T2_MSK_t2)	No DIF present anymore				Successful split, Testlet1 from this analysis can serve as an anchor for the t-test and effect-size calculation

FIM™: Functional Independence Measure; MSK: musculoskeletal rehabilitation; NEUR: neurological rehabilitation; t1: admission, t2: discharge, all: combination of time-points and rehabilitation-groups.

Combined effect size for repeated measures

	2-testlet Split2	
Mean person location	1.025	1.013
SD person location	2.3	2.341
Correlation of means	0.999	
Effect size	0.119	

SD: standard deviation.