Archives of Physical Medicine and Rehabilitation

journal homepage: www.archives-pmr.org

Archives of Physical Medicine and Rehabilitation 2013;94:527-35

ORIGINAL ARTICLE

Can We Scientifically and Reliably Measure the Level of Consciousness in Vegetative and Minimally Conscious States? Rasch Analysis of the Coma Recovery Scale-Revised

Fabio La Porta, MD,[a,b] Serena Caselli, PT,[a] Aladar Bruno Ianes, MD,[c] Olivia Cameli, MD,[d] Mario Lino, MD,[e] Roberto Piperno, MD,[d] Antonella Sighinolfi, MD,[e] Francesco Lombardi, MD,[f] Alan Tennant, PhD[g]

From the [a]Rehabilitation Medicine Unit, Azienda Unità Sanitaria Locale Modena, Modena, Italy; [b]PhD School in Advanced Sciences in Rehabilitation Medicine and Sports, Tor Vergata University, Rome, Italy; [c]Medical Direction, Segesta SpA, Korian Group, Milan, Italy; [d]"Casa dei Risvegli 'Luca De Nigris'" Hospital, Bologna, Italy; [e]"Villa delle Terme" Hospital, Segesta S.p.A., Korian Group, Florence, Italy; [f]Severe Brain Injury Unit, Azienda Unità Sanitaria Locale Reggio Emilia, Reggio Emilia, Italy; and [g]Department of Rehabilitation Medicine, Faculty of Medicine and Health, University of Leeds, Leeds, UK.

Abstract

Objectives: (1) To appraise, by the means of Rasch analysis, the internal validity and reliability of the Coma Recovery Scale-Revised (CRS-R) in a sample of patients with disorder of consciousness (DOC); and (2) to provide information about the comparability of CRS-R scores across persons with DOC across different settings and groups, including different etiologies.

Design: Multicenter observational prospective study.

Setting: Two rehabilitation wards, 1 intermediate care facility, and 2 nursing homes in Italy.

Participants: Consecutively admitted patients (N=129) for which assessments at 2 different time points were available, giving a total sample of 258 observations.

Interventions: Not applicable.

Main Outcome Measure: CRS-R.

Results: After controlling for any possible dependency between persons' measures collected at different time points, and for uniform differential item functioning by etiology showed by the visual subscale, Rasch analysis demonstrated adequate satisfaction of all the model's requirements, including adequate ordering of scoring categories, unidimensionality, local independence, invariance ($\chi^2_{21}=27.798$, $P=.146$), and absence of differential item functioning across patients' sex, age, time, and setting. The reliability (person separation index=.896) was adequate for individual person measurement. We devised a practical raw score to measure conversion tables based on the CRS-R calibrations.

Conclusions: The CRS-R is a psychometrically sound and robust measurement tool. The linear measures of ability derived from the CRS-R total scores do satisfy all the principles of scientific measurement and are sufficiently reliable for high stakes assessments, such as the diagnosis of the level of consciousness in individual patients. Future studies are needed to directly explore the capabilities of the CRS-R measures to reduce the risk of vegetative state misdiagnosis.

Archives of Physical Medicine and Rehabilitation 2013;94:527-35

The Coma Recovery Scale-Revised (CRS-R) was proposed by Giacino et al[1] as a bedside standardized neurobehavioral assessment tool incorporating the current diagnostic criteria for vegetative state (VS), minimally conscious state (MCS), and emergence from the MCS.[2] It consists of 29 hierarchically organized items grouped into 6 subscales addressing auditory, visual, motor, oromotor/verbal,

Presented to the Congress of the European Society of Physical and Rehabilitation Medicine, May 23–27, 2010, Venice, Italy; and the Congress of the European Federation for Research in Rehabilitation, May 26–28, 2011, Riva del Garda, Italy.

No commercial party having a direct financial interest in the results of the research supporting this article has or will confer a benefit on the authors or on any organization with which the authors are associated.

0003-9993/13/$36 - see front matter © 2013 by the American Congress of Rehabilitation Medicine

http://dx.doi.org/10.1016/j.apmr.2012.09.035

communication, and arousal functions.[1] The first 5 CRS-R subscales provide ordered score categories that are either linked to diagnoses of VS, MCS, or emergence from the MCS thus operationalizing the diagnostic criteria for these conditions. The total score, generated by summing together the subscale scores, may be used to track the changes of level of consciousness (LOC) over time,[3] although it is of limited diagnostic utility, because it is not linked to any diagnostic criteria. It is believed that the CRS-R may improve the accuracy of the differential diagnosis among individuals with disorder of consciousness (DOC), thus contributing to the reduction of the very high misdiagnosis rates (up to 37%−43%)[4-6] associated with a false positive diagnosis of VS.[7]

Several studies have assessed the reliability of the CRS-R under the classical test theory framework.[1,8-12] As shown in supplemental table S1, available online only at the Archives website: www.archives-pmr.org, several studies indicate that the various reliability coefficients for the single CRS-R subscales fell below the minimum recommended value for individual person measurement (.850) and, in some instances, even below the minimum value for group measurement (.700).[13] These findings may be explained considering that single-item scales (ie, the individual CRS-R subscales) are prone to large measurement errors leading to low reliability.[14] Such susceptibility to measurement error may compromise the intended use of the CRS-R as a high stakes tool for the diagnosis of LOC aiming at reducing VS misdiagnosis in individual patients.

On the other hand, it is well known that summative rating scales are generally more reliable than single-item scales, because the unavoidable random errors associated with the ratings of each item would cancel out if items were summed together to give a total score.[15,16] The latter, used together with the CRS-R subscales, may increase the reliability of the tool and, hence, reduce the risk of misdiagnosis. However, before considering this possibility, there must be evidence that summing together the CRS-R subscales to generate a total score is a legitimate procedure.[17] This evidence may be sought with new psychometric methods, such as Rasch analysis, which supplement validity and reliability data provided by the classical test theory approach. Rasch analysis is the process of iteratively testing whether the data meet the assumptions of the Rasch model (a mathematical model based on the work of the Danish mathematician Georg Rasch), which is known to operationalize the formal axioms of additive conjoint measurement.[18] Adequate fit to this model implies not only the legitimacy of summing the item scores to generate a total score, but the latter can also be transformed into an interval scale, whose unit of measurement is the logit[19]. In view of the item-free and sample-free calibration properties, as well as of the lack of distributional assumptions of the model, this interval scale can satisfy all the principles of scientific measurement,[18,20,21] thus allowing the comparability of measures across subjects and samples.[21]

List of abbreviations:

CRS-R	**Coma Recovery Scale-Revised**
DIF	**differential item functioning**
DOC	**disorder of consciousness**
GOS	**Glasgow Outcome Scale**
LOC	**level of consciousness**
MCS	**minimally conscious state**
NH	**nursing home**
PSI	**person separation index**
VS	**vegetative state**

Thus, for the current study, our goal was to fully appraise the internal construct validity (including the invariance of CRS-R totals scores across different etiologies and settings) and reliability of the CRS-R within the framework of Rasch modeling.

Methods

Participants, setting, and instruments

Data were collected prospectively across 5 different Italian facilities, including 2 rehabilitation wards, 1 intermediate care facility, and 2 nursing homes (NHs), between July 2009 and March 2012. All patients aged 18 to 75 years with a diagnosis of DOC as a result of an acquired etiology admitted to these units were included in this study. Exclusion criteria were preexisting neurologic degenerative pathologies and/or concurrent illnesses (eg, cancer) likely to affect survival within 6 months. Medically unstable patients were also temporarily excluded until their condition had improved sufficiently.

Data collection, based on the Italian version of the CRS-R, was performed by 12 raters who were all experienced in the care of this patient group, although their experience in using the CRS-R was variable, ranging from 2 months to 3 years. All raters used the developers' written scoring guidelines in order to minimize interrater variability.[3] All patients were assessed twice: first at enrollment, and then again at follow-up, after completion of the rehabilitation program or, for the remaining patients, after about 3 months. After each CRS-R assessment, the Glasgow Outcome Scale (GOS) and the Disability Rating Scale were also administered for external validation purposes.

Legal representatives of the incapacitated patients gave their informed consent for enrollment in the study, which was undertaken in compliance with the ethical principles set forth in the Helsinki Declaration.[22]

Rasch analysis

The Rasch model's assumptions and the Rasch analysis procedures (here based under the partial credit parameterization of the model) have been described in detail elsewhere.[19,23-27] We also assessed the reliability of the CRS-R (ie, its precision) using a person separation index (PSI), which provides estimates of the internal consistency reliability equivalent to Cronbach alpha.[24]

Specific analytical strategies: assessing and dealing with repeated measures and differential item functioning

Within the current study, specific analytical strategies (outlined in fig 1) were devised to deal with repeated measures and differential item functioning (DIF).

We controlled for any possible dependency between persons' measures collected at different time points following the strategy outlined in the sections A to D of figure 1. The assessment of the impact of repeated measures was performed by comparing the person ability estimates provided by an unconstrained Rasch analysis on the whole sample and by a constrained analysis according to the procedure suggested by Mallinson,[28] described in detail elsewhere.[26]

After obtaining a final solution, which satisfied the model's stochastic assumptions, we performed a DIF analysis[23,24] in order to assess the invariance of the item hierarchy across relevant group factors, such as sex, age, etiology, enrollment facility, time since lesion, and between assessments.[25,26] An item is said to display DIF (or item bias) if it gives different success rates for 2 or more groups at the same ability level.[29] Two kinds of DIF can be identified[23,24]: uniform DIF and nonuniform DIF, if the item bias remains constant or varies across all ability levels. If an item is affected by nonuniform DIF, it should be deleted as such violation of group invariance cannot be corrected. Whereas in cases of uniform DIF, it is either possible to delete the item or to split it by group level, which allows the item difficulty to vary across the various level of the person factor showing DIF.[23] Considering that sometimes there might be several instances of DIF affecting several items at one time and that any subsequent item deletion would change the original scale structure, should we find items affected by DIF, we would assess its real impact on person estimates by following the strategy outlined schematically in the sections E to G of figure 1. Specifically, we would do so by comparing the person estimates provided by 2 Rasch analyses based on the set with DIF and on a purified set (ie, without DIF), according to the strategy suggested by Tennant and Pallant.[30]

The impact of repeated measures and DIF would be considered negligible should the differences between each pair of person estimates generated from the comparisons previously outlined be less than .50 logits.[31]

Statistical notes, software, and sample size issues

We used SPSS[a] for descriptive statistics, whereas we carried out the Rasch analysis using the RUMM2030 software.[b] We estimated that a sample size of 250 observations would be sufficient to estimate item difficulty, with $\alpha = .01$ to $< \pm .50$ logits, irrespective of the targeting of persons to the items.[32] Throughout, we used a significance value of .05 adjusted for the number of tests by Bonferroni correction.[33]

Results

Participants recruited and scale statistics

All observations were collected on a convenience sample of 129 patients for whom both enrollment and follow-up assessments were available, thus making a total sample of 258 observations available for the analyses. Sample descriptive statistics are summarized in table 1. The median total CRS-R score for the whole observation sample was 7 (range, 0−23; mean ± SD, 8.7±5.3), and all 24 CRS-R scale scores were represented.

Rasch analysis

Unique assessments sample (N=129)

The first Rasch analysis (see fig 1A and table 2, analysis 1) showed adequate fit to the model. Particularly, all subscales had an ordered structure in terms of response categories and fit the model individually. The scale was strictly unidimensional, and there was no significant local dependency between items. Overall, the data fit the model well ($\chi^2_{12} = 6.200$, $P = .906$), and the reliability was compatible with measurements at the individual level (PSI=.886, $\alpha = .859$).

Dealing with repeated measures (N=258)

The unconstrained Rasch analysis on the whole 258 observation sample (see fig 1B and table 2, analysis 2) confirmed adequate fit to the Rasch model ($\chi^2_{18} = 14.681$, $P = .684$). These findings were confirmed by the subsequent anchored analysis (see fig 1C and table 2, analysis 3), where the exported item difficulty estimates as well as the thresholds from the unique assessments analysis were anchored to the whole observation sample. Particularly, there was evidence of adequate model fit ($\chi^2_{18} = 22.380$, $P = .216$) and reliability (PSI=.897, $\alpha = .888$). At this stage, the comparison between the person estimates provided by the unconstrained and the anchored analysis (see fig 1D) showed no differences >.50 logits for any individual comparisons (mean absolute logit difference=.136), suggesting the lack of any significant effect of the repeated-measures design on person estimates.

Assessing of and dealing with DIF (N=258)

At this stage, we also performed a DIF analysis on the anchored set (see fig 1E) by testing the following factors: sex, age (≤47y, ≥48y), etiology (hemorrhage, traumatic brain injury, anoxic, and other etiologies), time duration since lesion (≤142d, 143−419d, ≥420d), typology of assessment (enrollment, follow-up), time distance between assessments (<90d, ≥90d), and tipology of setting (NH and intermediate care facilities, rehabilitation centers). The DIF analysis showed the presence of uniform DIF by etiology for the visual subscale. Particularly, groups of persons with anoxic brain injury found this item systematically more difficult than groups of persons of equal ability with brain injury because of the other etiologies ($F_2 = 8.426$; $P < .000$). Also, the DIF analysis showed the presence of nonuniform DIF for the oromotor subscale, both by etiology ($F_6 = 3.653$, $P = .002$) and setting ($F_3 = 9.034$, $P < .000$).

As detailed in sections F to G of figure 1, after elimination of the biased items, the comparison of the person estimates derived from the pure set and the anchored full set showed that 7% of those estimates differed by >.50 logits, thus suggesting a significant impact of uniform DIF by etiology on person estimates. As a consequence, we split the visual subscale by etiology, allowing a separate estimate of item difficulty for the anoxic brain injury group and the hemorrhagic and traumatic brain injury groups. After controlling for any eventual person dependency because of repeated measures by anchoring the new item set to the item estimates derived from the unique assessments analysis, the oromotor subscale still showed nonuniform DIF by setting ($F_3 = 8.791$, $P < .000$). However, by repeating the steps F to G of figure 1 for this item, we were able to demonstrate the lack of differences >.50 logits across the person estimates derived from the pure set and the anchored full set, suggesting the lack of a significant impact of nonuniform DIF for the oromotor subscale.

Final solution (N=258)

The final solution (see table 2, analysis 4) thus showed adequate fit to the Rasch model ($\chi^2_{21} = 27.798$, $P = .146$). The scale was strictly unidimensional (proportion of significant t test=5%; binomial confidence interval for proportions, 2.4%−7.7%), and there was no significant item local dependency. All items showed ordered response categories and fit the model individually (table 3). The targeting graph of the CRS-R (fig 2) showed that persons were evenly spread across 10 logits, with negligible floor (0.4%) and ceiling effects (1.6%). The mean person ability of −1.161 logits indicated that the ability of the sample was slightly lower than the average difficulty of the CRS-R, set by default to 0 logits. The person reliability was adequate for individual person measurement (PSI=.896, $\alpha = .887$).[13] Given the PSI, persons could be separated in 4.3 strata, that is, the statistically distinct levels of ability that the CRS-R was able to reliably distinguish in this sample.[34] Because no rescoring nor item deletions were undertaken, the original CRS-R total score remained unchanged, ranging from 0 to 23. The item hierarchy (see table 3) was consistent with clinical expectations, because the

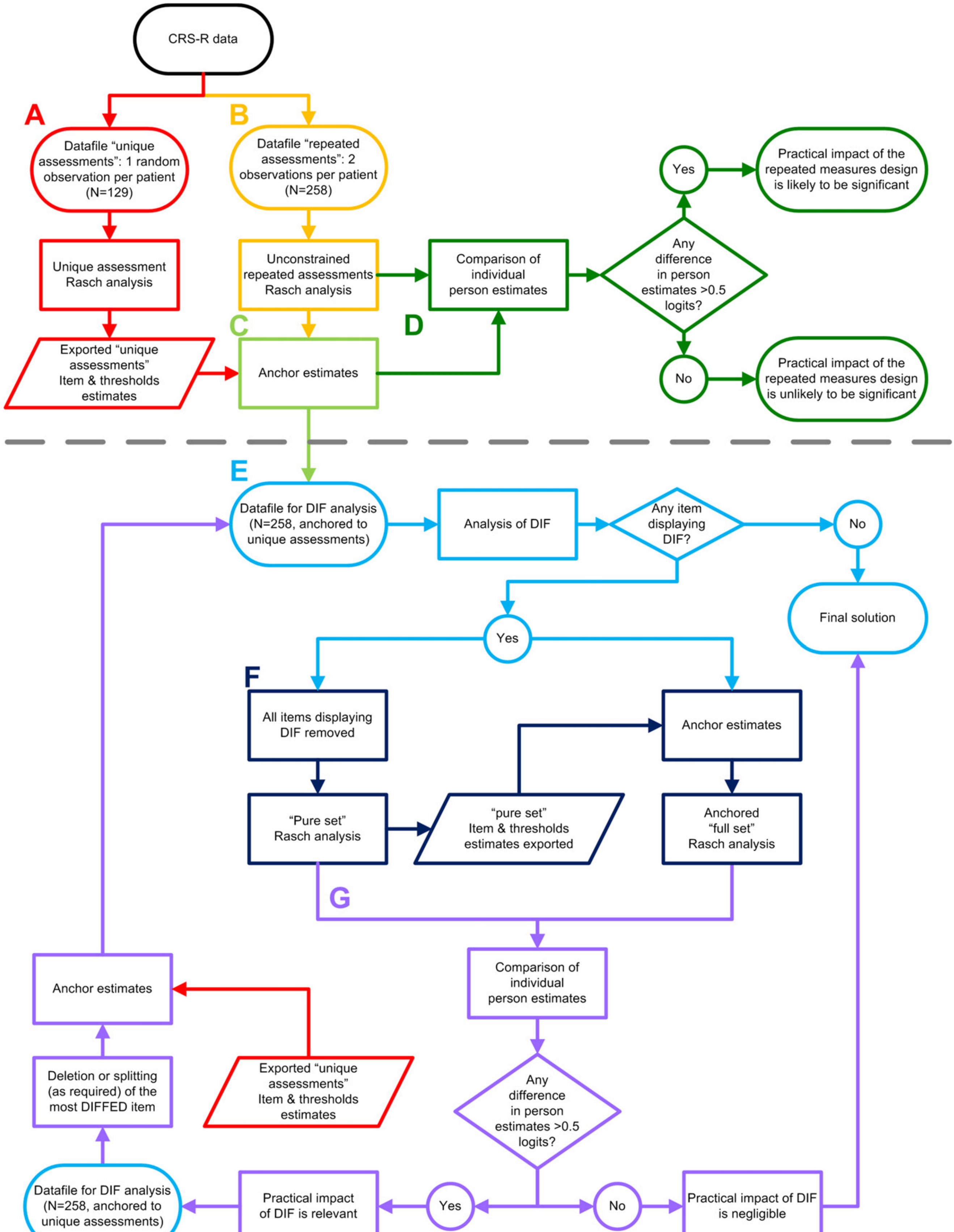

Fig 1 Outline of the analytical strategies adopted to deal with repeated measures and DIF. Sections A—D show the strategy adopted to control for any person dependency because of repeated measurements, whereas sections E—G deal with the strategy employed to account for DIF. For each

Table 1 Sample descriptive statistics

	Enrollment Assessment Only (n=129)				Enrollment and Follow-Up Assessments (N=258)	
	n	%	Mean ± SD	Median	n	%
Setting						
Rehabilitation	63	48.8				
Intermediate care facility	36	27.9				
NH	30	23.3				
Age (y)	129		47±20	48.6		
Sex						
Male	82	63.6				
Female	47	36.4				
Etiology						
Traumatic brain injury	57	44.2				
Hemorrhagic stroke	34	26.4				
Anoxic brain injury	27	20.9				
Ischemic stroke	6	4.7				
Other etiologies*	5	3.9				
Time since lesion (d)						
Whole sample	129	100.0	421±599	162		
Rehabilitation	63	48.8	119±102	88		
Intermediate care facility	36	27.9	624±564	434		
NH	30	23.3	920±923	570		
Diagnosis (GOS)						
VS	103	79.8			184	71.3
Severe disability	24	18.6			57	22.1
Moderate disability	2	1.6			18	6.6

* Examples include meningoencephalitis and poisoning.

easiest subscales were the arousal and the motor subscales, whereas the most difficult item was the communication subscale.

On the basis of the item calibrations, it was possible to construct 2 tables to convert raw scores to measures (table 4) for individuals with traumatic or hemorrhagic brain injury (where the visual subscale was just third in the item difficulty hierarchy) and for persons with anoxic brain injury and other etiologies (where the visual subscale was the penultimate most difficult item).

Discussion

To our knowledge, this is the first published study that fully appraised the internal validity and reliability of the CRS-R on a sample of patients with DOC within the framework of Rasch analysis. Our results suggest that the presence of DIF by etiology for the visual subscale prevented invariance of the measures across the etiology of the DOC. However, after adjusting for this item bias, the CRS-R demonstrated excellent internal construct validity,[35] thus enabling us to transform its total scores into linear measures of ability that satisfied all the principles of scientific measurement,[18,35] and were also sufficiently reliable for individual patient measurement.[13]

The final Rasch analysis was based on a 258 observations sample including repeated measures at 2 time points for all patients. In order to control for any possible time series dependency, we followed the procedure suggested by Mallinson,[28] which allowed us to measure persons at different time points

patient (step A), we randomly selected either the enrollment or the follow-up assessment. These unique observations were subjected to a first Rasch analysis (unique assessments) and the corresponding item and threshold estimates were exported. Subsequently, we performed a Rasch analysis on the whole sample (step B) without applying any constraint (unconstrained repeated assessments), thus without controlling for any eventual person dependency across different time points. After this, we performed a further Rasch analysis (step C) by anchoring the unique assessment item difficulties and Rasch-Andrich thresholds estimates (exported at step A) to the whole observation sample (anchored repeated assessments). Finally, in order to assess the impact of repeated measures on ability estimates, we compared the person estimates derived from the unconstrained analysis with those produced by the anchored analysis (step D). A DIF analysis (step E) was conducted within the context of the anchored repeated assessment Rasch analysis (step C). Should 1 or more item show evidence of DIF, we adopted the following strategy, suggested by Tennant and Pallant[30] (step F): (1) we removed all items affected by a statistically significant DIF from the full item set; (2) from the obtained set (pure set), we exported the item parameter estimates for the 3 items displaying the least DIF; and (3) we anchored the exported item parameter estimates to the full set (anchored full set), and therefore the person ability estimates were based on the same measurement scale defined by the pure item showing the least DIF. Finally (step G), we compared the person estimates from the pure and the anchored sets in a spreadsheet in order to assess the practical impact of DIF. Thus, should we find DIF for any item, we would employ this strategy in order to assess its impact on person estimates. Only in the case of significant impact of DIF on person estimates (as defined in the figure and in the Methods section), we would adjust for DIF by item splitting or deletion, as necessary. Finally, considering the need of controlling for any possible person dependency across time, after each item splitting or deletion, we would anchor the item estimates to those provided by the unique assessments analysis. Abbreviation: DIFFED, affected by DIF.

Table 2 Overall fit to the Rasch model for the CRS-R

Analysis			Item Residual	Person Residual	Item-Trait Interaction		Reliability		Unidimensionality *t* Test	
No.	Description	n	Mean ± SD	Mean ± SD	χ^2 (df)	*P*	PSI	α	PST (%)	BCI (%)
1	Unique assessment sample	129	−0.276±0.422	−0.362±0.873	6.200 (12)	.906	.886	.859	3.9	0.1−7.6
2	Repeated assessments sample, unconstrained	258	−0.416±0.813	−0.415±0.906	14.681 (18)	.684	.889	.876	4.3	1.6−6.9
3	Repeated assessments sample, anchored to analysis 1	258	−0.297±0.859	−0.401±0.932	22.380 (18)	.216	.897	.888	5.0	2.4−7.7
4	Repeated assessments sample, anchored to analysis 1, visual subscale split by etiology	258	−0.243±0.803	−0.388±0.921	27.798 (21)	.146	.896	.887	5.0	2.4−7.7
	Recommended values	NA	0.000±1.000	0.000±1.000	NA	>.006*	>.850†	>.850†	<5.0‡	Lower BCI <5‡

Abbreviations: BCI, binomial confidence interval for PST; NA, not applicable; *P*, Bonferroni-corrected chi-square value; PST, proportion of significant *t* test carried out on the estimates that, within a principal component analysis of residuals, loaded positively and negatively (factor loading >±.30) on the first component.

* Bonferroni-corrected value of .05, indicative of statistical significance, will vary by analysis; this value is referred to the final solution.

† Value of >.850 indicates precision of measurement also at the individual level, whereas a value of >.700 indicates precision only at the group level.

‡ Strict unidimensionality is considered achieved either when PST is <5% or, alternatively, when the lower bound of its BCI is <5%.

within the same frame of measurement.[26] Although we were able to demonstrate the lack of any substantial time-series dependency effect on the person estimates of the unanchored analysis, we preferred to use the estimates from the anchored analysis in order to avoid possible hidden violations of the assumption of statistical independence among the observations at the item level. Following the previously mentioned procedure, we were able to employ a sample that, to our knowledge, was the largest employed to date in a CRS-R validation study. Unlike previous reports[1,10] where underused item score categories had been reported and not all available total scores had been affirmed, in our study, all CRS-R total scores were represented with minimal floor and/or ceiling effects. Because the sample was adequately targeted, it included patients with the full spectrum of LOC abilities measured by the CRS-R, ranging from VS to emergence from the MCS. This suggests that the sharp prevalence of VS patients in the sample (73.1%), as suggested by the GOS, may reflect some misdiagnosis. The latter is also likely considering how single items scales, such as the GOS, are prone to measurement error.[14]

Table 3 Item parameters and fit statistics for the CRS-R (n=258, analysis number 4)

CRS-R Subscales	Location	SE	Fit Residual	χ^2	*P**
CRS6—arousal	−1.914	.124	−0.036	4.898	.179
CRS3—motor	−0.205	.082	0.641	3.492	.322
CRS2b—visual (TBI-hemorrhage)	−0.097	.100	−0.207	1.692	.639
CRS1—auditory	−0.064	.102	−1.057	4.954	.175
CRS4—oromotor	0.027	.117	0.903	4.657	.199
CRS2a—visual (anoxic-ischemic)	0.164	.171	−0.964	4.093	.252
CRS5—communication	2.186	.177	−0.984	4.013	.260

NOTE. CRS-R items are ordered by progressively increasing difficulty from top to bottom. The location is expressed in logits. As the visual subscale was split for etiology, both etiology-specific versions were reported. The degrees of freedom for each chi-square were 3 for all items.

Abbreviations: *P*, chi-square probability; TBI, traumatic brain injury.

* Bonferroni-corrected *P* indicating statistical significance at the .05 level was .006.

The enlargement of the sample allowed the emergence of a significant issue, that is, DIF. This was handled with a conservative strategy aimed at minimizing the changes to the scale structure. For instance, this strategy demonstrated the lack of significant impact on the estimates of the nonuniform DIF displayed by the oromotor subscale, thus avoiding the deletion of this item. On the other hand, the impact on person estimates of the uniform DIF by etiology displayed by the visual subscale was not only statistically significant, but was also consistent with clinical expectations. Traumatic and hemorrhagic brain injuries are less likely to cause the selective damage to visual pathways that is known to occur more frequently in anoxic brain injury as a consequence of the selective cortical necrosis typical of this condition.[36] We could deal with this bias with a conservative approach based on splitting the visual subscale by etiology. In this way we were able to adjust the person estimates for this factor that, if not accounted for, would have prevented the invariance of the scale across different causes of severe brain injury.

This study provided strong evidence for the internal construct validity[35] of the CRS-R. The fact that the very strict measurement criteria of the Rasch model in terms of unidimensionality, local independence, and invariance at the item and total score level were satisfied without any significant modification to the scale structure indicates that the CRS-R is a psychometrically sound and very robust measure based on excellent item design. This is also suggested by the stability of the ordering of the score categories for the CRS-R subscales across different settings and raters with variable experience in using the tool. The psychometric stability and robustness of the CRS-R may be explained

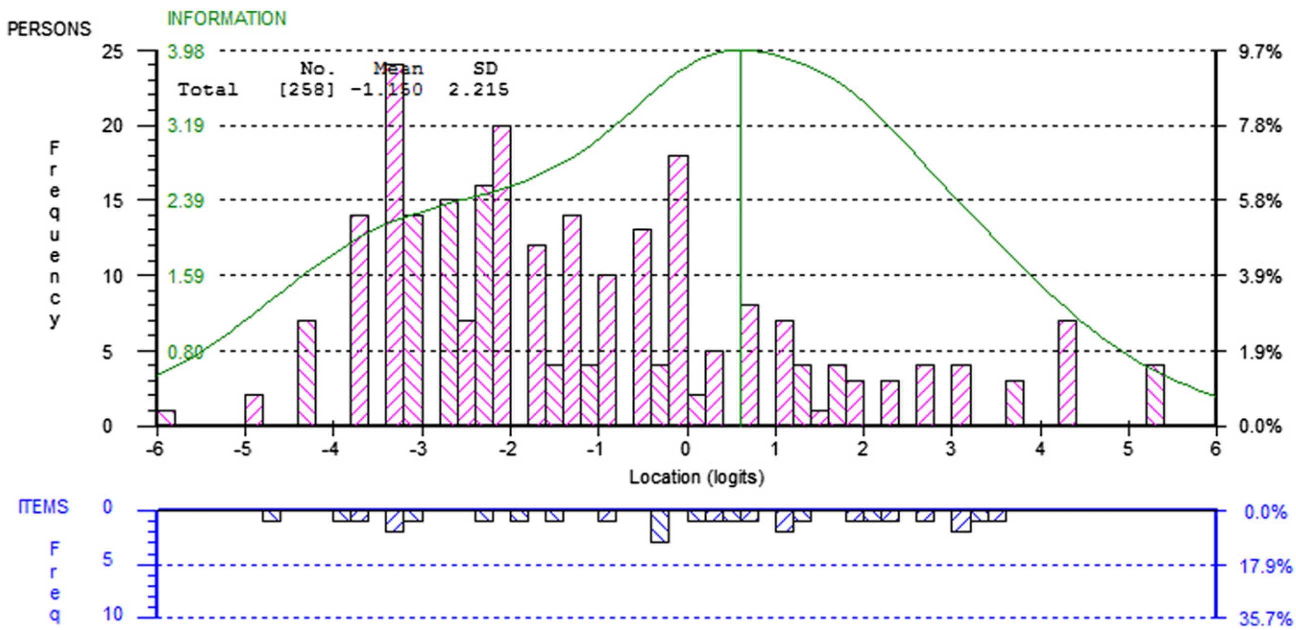

Fig 2 Targeting of the CRS-R (n=258). Observations (n=258) and subscale thresholds are displayed, respectively, in the upper and the lower part of the graph, separated by the logit scale. Grouping set to interval length of 0.20 making 60 groups. The scores provided by both the etiology-specific versions of the scale were used. Abbreviation: Freq, frequency.

Table 4 Raw score to measure estimates conversion table for the CRS-R based on the original sample calibrations

	Traumatic and Hemorrhagic Brain Injury				Anoxic and Other Causes of Brain Injury			
Raw Score	Logit Scale	±95%CI	0–100 Scale	±95%CI	Logit Scale	±95%CI	0–100 Scale	±95%CI
0	−5.871	1.335	0.0	23.6	−5.823	1.364	0.4	24.1
1	−4.992	0.974	7.9	17.2	−4.941	0.986	8.4	17.4
2	−4.296	0.817	14.2	14.4	−4.228	0.833	14.8	14.7
3	−3.744	0.748	19.2	13.2	−3.650	0.771	20.0	13.6
4	−3.242	0.718	23.7	12.7	−3.111	0.746	24.9	13.2
5	−2.753	0.704	28.1	12.4	−2.571	0.736	29.7	13.0
6	−2.264	0.695	32.5	12.3	−2.024	0.724	34.6	12.8
7	−1.781	0.684	36.8	12.1	−1.490	0.698	39.5	12.3
8	−1.316	0.669	41.0	11.8	−1.005	0.659	43.8	11.6
9	−0.876	0.653	45.0	11.5	−0.591	0.617	47.6	10.9
10	−0.460	0.636	48.7	11.2	−0.239	0.585	50.7	10.3
11	−0.065	0.620	52.3	10.9	0.074	0.564	53.5	10.0
12	0.310	0.604	55.7	10.7	0.369	0.553	56.2	9.8
13	0.666	0.591	58.9	10.4	0.661	0.549	58.8	9.7
14	1.003	0.579	61.9	10.2	0.956	0.550	61.5	9.7
15	1.323	0.573	64.8	10.1	1.255	0.554	64.2	9.8
16	1.634	0.573	67.6	10.1	1.557	0.562	66.9	9.9
17	1.950	0.584	70.4	10.3	1.868	0.576	69.7	10.2
18	2.286	0.607	73.5	10.7	2.198	0.601	72.7	10.6
19	2.661	0.644	76.8	11.4	2.567	0.639	76.0	11.3
20	3.095	0.700	80.8	12.4	2.997	0.697	79.9	12.3
21	3.620	0.789	85.5	13.9	3.521	0.789	84.6	13.9
22	4.314	0.965	91.7	17.0	4.222	0.968	90.9	17.1
23	5.232	1.313	100.0	23.2	5.157	1.318	99.3	23.3

NOTE. As the visual subscale was split for etiology, both etiology-specific person estimates were reported. The latter are expressed both in logits and in a 0 to 100 (or percentage) scale.
Abbreviation: CI, confidence interval (equal to 1.96 standard error of measurement).

considering its conceptually rigorous construction process,[1,2] including the fact that the current scale was derived from a refinement of a previous version (published in 1991)[37] on the basis of clinical experience and a Rasch analysis.[1] Our study also demonstrated that the CRS-R provides invariant and comparable measures irrespective of the temporal evolution of the underlying condition, setting, and age and sex of the patients. On the other hand, the adjustment for DIF by etiology of the visual subscale allowed comparability of the CRS-R measures irrespective of the cause of the DOC.

The CRS-R measures hold a person reliability above the minimum recommended criterion (0.850) for measurement at the individual level,[13] suggesting that those may be a reliable adjunctive diagnostic tool in high stakes situations as the diagnosis of LOC in individual patients. Despite the validity and reliability of the CRS-R measures, it should be borne in mind, though, that the diagnosis of LOC is a complex task requiring the careful consideration of other factors, such as unpredictable fluctuations of the arousal level, positioning, associated sensorial, motor and cognitive impairments, level of medical stability, and medications administered.[2]

Study limitations

Because VS and MCS are rare conditions,[38,39] it may be difficult to collect a large enough and well-targeted sample to obtain stable item calibrations.[32] Although this problem was overcome by allowing repeated observations, the sample was not large enough to allow proper confirmation of the model fit with a revalidation sample, which would have further minimized the risk of capitalizing on chance with respect to fit to the model. Given this limitation, these findings will require replication in the context of a larger multicenter study aimed at confirming the fit to the model and the stability of the raw score to measure change tables for the CRS-R.

Conclusions

The CRS-R is a psychometrically sound and robust measurement tool for patients with DOC, with adequate internal construct validity and reliability under the Rasch analysis framework. This allowed the construction of 2 tables to convert raw scores into measures that are simple methods to transform the CRS-R total scores into linear estimates of ability. Because these satisfy the requirements for interval-level measurement, clinicians and researchers may want to use these scientific measures rather than total scores for the possibility of using parametric statistics (eg, analysis of variance).[24,40] Furthermore, because the correct diagnosis of LOC requires repeated assessment over time,[2] the use of the CRS-R measures may also be very helpful for the correct interpretation of change-scores.[24] Future studies are needed to directly explore the capabilities of the CRS-R measures to reduce the VS misdiagnosis rate.

Suppliers

a. SPSS version 13; SPSS Inc, 233 S Wacker Dr, 11th Fl, Chicago, IL 60606.
b. RUMM2030 professional edition, version 5.4; RUMM Laboratory Pty Ltd, 14 Dodonaea Ct, Duncraig, WA, Australia 6023.

Keywords

Consciousness disorders; Outcome assessment (health care); Persistent vegetative state; Psychometrics; Rehabilitation

Corresponding author

Fabio La Porta, MD, Unità Operativa di Medicina Riabilitativa, Nuovo Ospedale Civile "S. Agostino-Estense," via Giardini 1455, 41126, Modena, Italy. *E-mail address:* fabiolaporta@mail.com.

Acknowledgments

We thank the following professionals working with Segesta S.p.A. (Korian Group) for data collection: Annamaria D'Avanzo, MD, Marco Breschi, MD, Illy Dine, MD, Paola Pacini, PT, Andrea Bracali, PT, Anna Puccetti, PT, and Antonella Fedi, PT.

References

1. Giacino JT, Kalmar K, Whyte J. The JFK Coma Recovery Scale-Revised: measurement characteristics and diagnostic utility. Arch Phys Med Rehabil 2004;85:2020-9.
2. Seel RT, Sherer M, Whyte J, et al. Assessment scales for disorders of consciousness: evidence-based recommendations for clinical practice and research. Arch Phys Med Rehabil 2010;91: 1795-813.
3. Lombardi F, Gatta G, Sacco S, Muratori A, Carolei A. The Italian version of the Coma Recovery Scale-Revised (CRS-R). Funct Neurol 2007;22:47-61.
4. Andrews K, Murphy L, Munday R, Littlewood C. Misdiagnosis of the vegetative state: retrospective study in a rehabilitation unit. BMJ 1996; 313:13-6.
5. Childs NL, Mercer WN. Misdiagnosing the persistent vegetative state. Misdiagnosis certainly occurs. BMJ 1996;313:944.
6. Schnakers C, Vanhaudenhuyse A, Giacino J, et al. Diagnostic accuracy of the vegetative and minimally conscious state: clinical consensus versus standardized neurobehavioral assessment. BMC Neurol 2009;9:35.
7. Kalmar K, Giacino JT. The JFK Coma Recovery Scale—Revised. Neuropsychol Rehabil 2005;15:454-60.
8. Schnakers C, Majerus S, Giacino J, et al. A French validation study of the Coma Recovery Scale-Revised (CRS-R). Brain Inj 2008;22: 786-92.
9. Lovstad M, Froslie KF, Giacino JT, Skandsen T, Anke A, Schanke AK. Reliability and diagnostic characteristics of the JFK coma recovery scale-revised: exploring the influence of rater's level of experience. J Head Trauma Rehabil 2010;25:349-56.
10. Sacco S, Altobelli E, Pistarini C, Cerone D, Cazzulani B, Carolei A. Validation of the Italian version of the Coma Recovery Scale-Revised (CRS-R). Brain Inj 2011;25:488-95.
11. Simoes JF, Jesus LM, Voegeli D, Sa-Couto P, Fernandes J, Morgado M. Assessment of comatose patients: a Portuguese instrument based on the Coma Recovery Scale - revised and using nursing standard terminology. J Adv Nurs 2011;67:1129-41.
12. Noé E, Olaya J, Navarro MD, et al. Behavioral recovery in disorders of consciousness: a prospective study with the Spanish version of the Coma Recovery Scale-Revised. Arch Phys Med Rehabil 2012;93:428-433.e12.
13. Reeve BB, Hays RD, Bjorner JB, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Med Care 2007;45(5 Suppl 1):S22-31.
14. Nunnally JC. Psychometric theory. New York: McGraw-Hill; 1994.

15. Hobart J, Cano S, Zajicek J, Thompson A. Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. Lancet Neurol 2007;6:1094-5.
16. Hobart J. Measuring outcomes in clinical trials of stroke: time for state-of-the-art clinical trials to reject state-of-the-ark rating scales. J Neurol 2007;254:1119.
17. Hobart J, Cano S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. 2009. Available at: http://www.hta.ac.uk/fullmono/mon1312.pdf. Accessed July 2, 2009.
18. Perline R, Wright BD, Wainer H. The Rasch model as additive conjoint measurement. Applied Psychological Measurement 1979;3: 237-55.
19. Andrich D. Rasch models for measurement. London: Sage Publications; 1988.
20. Bond TG, Fox CM. Applying the Rasch model: fundamental measurement in the human science. 2nd ed. Mahwah: Lawrence Erlbaum Associates Publishers; 2007.
21. Tesio L. Measuring behaviors and perceptions: Rasch analysis as a tool for rehabilitation research. J Rehabil Med 2003;35:105-15.
22. World Medical Association. WMA Declaration of Helsinki: ethical principles for medical research involving human subjects. 2008. Available at: http://www.wma.net/en/30publications/10policies/b3/. Accessed January 31, 2009.
23. Tennant A, Penta M, Tesio L, et al. Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model: the PRO-ESOR project. Med Care 2004;42(1 Suppl): I37-48.
24. 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: 1358-62.
25. La Porta F, Franceschini M, Caselli S, Cavallini P, Susassi S, Tennant A. Unified Balance Scale: an activity-based, bed to community, and aetiology-independent measure of balance calibrated with rasch analysis. J Rehabil Med 2011;43:435-44.
26. La Porta F, Caselli S, Susassi S, Cavallini P, Tennant A, Franceschini M. Is the Berg Balance Scale an internally valid and reliable measure of balance across different etiologies in neurorehabilitation? A revisited Rasch analysis study. Arch Phys Med Rehabil 2012;93:1209-16.
27. Panella L, La Porta F, Caselli S, Marchisio S, Tennant A. Predicting the need for institutional care shortly after admission to rehabilitation: Rasch analysis and predictive validity of the BRASS Index. Eur J Phys Rehabil Med 2012;48:443-54.
28. Mallinson T. Rasch analysis of repeated measures. Rasch Measure Trans 2011;25:1317. Available at: http://www.rasch.org/rmt/rmt251b.htm. Accessed October 1, 2011.
29. Holland PW, Wainer H. Differential item functioning. Hillsdale: Lawrence Erlbaum Associates; 1993.
30. Tennant A, Pallant J. DIF matters: a practical approach to test if differential item functioning makes a difference. Rasch Measure Trans 2007;20(4):1082-4.
31. Wright BD, Panchapakesan N. A procedure for sample-free item analysis. Educational and Psychological Measurement 1969;29:23-48.
32. Linacre JM. Sample size and item calibration [or person measure] stability. Available at: http://www.rasch.org/rmt/rmt74m.htm. Accessed February 25, 2006.
33. Bland J, Altman D. Multiple significance tests: the Bonferroni method. BMJ 1995;310:170.
34. Wright BD, Masters GN. Rating scale analysis. Chicago: MESA Pr; 1982.
35. Kucukdeveci AA, Tennant A, Grimby G, Franchignoni F. Strategies for assessment and outcome measurement in physical and rehabilitation medicine: an educational review. J Rehabil Med 2011;43:661-72.
36. Fitzgerald A, Aditya H, Prior A, McNeill E, Pentland B. Anoxic brain injury: clinical patterns and functional outcomes. A study of 93 cases. Brain Inj 2010;24:1311-23.
37. Giacino JT, Kezmarsky MA, DeLuca J, Cicerone KD. Monitoring rate of recovery to predict outcome in minimally responsive patients. Arch Phys Med Rehabil 1991;72:897-901.
38. Monti MM, Laureys S, Owen AM. The vegetative state. BMJ 2010; 341:c3765.
39. Wilson SL, Powell GE, Brock D, Thwaites H. Behavioural differences between patients who emerged from vegetative state and those who did not. Brain Inj 1996;10:509-16.
40. Svensson E. Guidelines to statistical evaluation of data from rating scales and questionnaire. J Rehabil Med 2001;33:47-8.

Supplemental Table 1 Summary of reliability study results of the CRS-R

	Giacino and Kalmar (2004)[7]			Schnakers et al (2008)[8]			Lovstad et al (2010)[9]			Simões et al (2011)[11]			Sacco et al (2011)[10]		
Study design and setting															
No. of centers	1			5			6			1			1		
Assessment setting	R			A, R, NH			R, NH			A			R		
Sample size	80			77			31			20			38		
No. of raters	2			24			8			2			2		
Sample size/rater ratio	40			3.2			3.8			10			19		
Reliability	IRR	TRT	ICR	IRR	TRT	ICR	IRR	TRT	ICR	IRR	TRT	ICR	IRR	TRT	ICR
Auditory subscale	$\kappa = .86$	$\kappa = 0.63$†		$\kappa = .82$*			$\kappa = .90$	$\kappa = .71$*		ICC = 0.99	ICC = .86		$\kappa_w = .65$†	$\kappa_w = 0.80$*	
Visual subscale	$\kappa = .58$†	$\kappa = 0.90$		$\kappa = .85$			$\kappa = .46$†	$\kappa = .86$		ICC = 1.00	ICC = .88		$\kappa_w = .71$*	$\kappa_w = 0.84$*	
Motor subscale	$\kappa = .78$*	$\kappa = 1.00$		$\kappa = .93$			$\kappa = .67$†	$\kappa = .73$*		ICC = 0.98	ICC = .81*		$\kappa_w = .79$*	$\kappa_w = 0.96$	
Oromotor subscale	$\kappa = .77$*	$\kappa = 0.23$†		$\kappa = .92$			$\kappa = .89$	$\kappa = .71$*		ICC = 0.96	ICC = .82*		$\kappa_w = .44$†	$\kappa_w = 0.85$	
Communication subscale	$\kappa = .88$	$\kappa = 0.89$		$\kappa = .98$			$\kappa = .62$†	$\kappa = .89$		ICC = 0.97	ICC = .82*		$\kappa_w = .88$	$\kappa_w = 0.88$	
Vigilance subscale	NA	NA		$\kappa = .74$*			NA	NA		ICC = 0.98	ICC = .84*		$\kappa_w = .51$†	$\kappa_w = 1.00$	
Total score	$\rho = .84$*	$\rho = 0.94$	$\alpha = .84$*	$\kappa = .80$*	NA	NA	$\kappa = .94$	NA	$\alpha = .74$*	ICC = 0.99	ICC = .87	NA	$\rho = .81$*	$\rho = 0.97$	$\alpha = .81$*

NOTE. Where several reliability values were available, we reported only the largest ones. Comparison of the various studies may be difficult in view of the fact that the classical psychometric properties reported are strictly sample-dependent and several reliability coefficients were used across different studies.

Abbreviations: α, Cronbach α; A, acute setting; CRS-R, CRS-R total score; ICC, intraclass correlation coefficient; ICR, internal consistency reliability; IRR, interrater reliability; κ, Cohen κ; κ_w, weighted κ; NA, not applicable; ρ, Spearman correlation coefficient; R, rehabilitation setting; TRT, test-retest reliability.

* Values compatible only with measurement at the group level ($\geq .70 < .85$), not at the individual level ($\geq .85$).

† Values not sufficiently reliable for any measurement ($< .70$), including measurement at the group level.