



HAL
open science

Determining the Interrater Reliability of the SOFMER Activity Score (version 2) for Individuals in Rehabilitation Centers

Lorraine Charvolin, Pascal Rippert, Sylvain Roche, Muriel Rabilloud, Marie-Doriane Morard, Julie Di Marco, Mickael Dinomais, Margaux Pouyfaucou, Rémi Gimat, Dominique Perennou, et al.

► To cite this version:

Lorraine Charvolin, Pascal Rippert, Sylvain Roche, Muriel Rabilloud, Marie-Doriane Morard, et al.. Determining the Interrater Reliability of the SOFMER Activity Score (version 2) for Individuals in Rehabilitation Centers. Archives of Physical Medicine and Rehabilitation, 2021, <10.1016/j.apmr.2021.11.005>. <hal-03582488>

HAL Id: hal-03582488

<https://hal.univ-grenoble-alpes.fr/hal-03582488v1>

Submitted on 22 Jul 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY-NC 4.0 - Attribution - Non-commercial use - International License

Determining the inter-rater reliability of the SOFMER Activity Score (version 2) for subjects in rehabilitation centers

(Running head: Inter-rater reliability of SOFMER Score)

Lorraine Charvolin ¹, Pascal Rippert ², Sylvain Roche ³, Muriel Rabilloud ³, Marie-Doriane Morard ¹, Julie Di Marco ⁴, Mickael Dinomais ⁵, Margaux Pouyfaucou ⁶, Rémi Gimat ⁷, Dominique Perennou ⁷, Laetitia Houx ⁸, Jean Iwaz ³, Gilles Rode ^{4,9}, Carole Vuillerot ¹⁰

¹ Service de Médecine Physique et de Réadaptation Pédiatrique (L'Escale), Hôpital Femme–Mère–Enfant, Hospices Civils de Lyon, Bron, France.

² Service Recherche et Épidémiologie Clinique, Pôle santé publique, Hospices Civils de Lyon, Lyon, France.

³ Université de Lyon, Lyon, France; Université Lyon 1, Villeurbanne, France; Hospices Civils de Lyon, Pôle Santé Publique, Service de Biostatistique-Bioinformatique, Lyon, France; CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Évolutive, Équipe Biostatistique-Santé, Villeurbanne, France.

⁴ Service de Médecine Physique et Réadaptation, Hôpital Henry-Gabrielle, Hospices Civils de Lyon, Saint-Genis-Laval, France.

⁵ Département de Médecine Physique et Rééducation, Centre Hospitalier Universitaire, Angers, France.

⁶ Service Médecine Physique et Rééducation Fonctionnelle, Centre Hospitalier Universitaire d'Angers, Angers, France; Centre de Rééducation et de Réadaptation Fonctionnelles Les Capucins, Angers, France; Service de Rééducation, Centre Hospitalier de Cholet, Cholet, France.

⁷ Service Rééducation Neurologique, Hôpital Sud Centre Hospitalier Universitaire de Grenoble-Alpes, Echirolles, France; Laboratoire de Psychologie et Neurocognition (LPNC), Université Grenoble-Alpes, Grenoble, France.

⁸ Service de Médecine Physique et de Réadaptation, Centre Hospitalier Régional et Universitaire de Brest, Brest, France; Inserm UMR 1101, Laboratoire de Traitement de l'Information Médicale (LaTIM), Brest, France; Service de Médecine Physique et de Réadaptation Pédiatrique, Fondation Ildys, Brest. France.

⁹ Integrative, Multisensory, Perception, Action and Cognition Team (IMPACT), Centre de Recherche en Neurosciences de Lyon (Inserm UMR-S, 1028, CNRS UMR 5292, Université Lyon 1, Université Saint-Etienne), Bron, France.

¹⁰ Service de Médecine Physique et de Réadaptation Pédiatrique (L'Escale), Hôpital Femme–Mère–Enfant, Hospices Civils de Lyon, Bron, France; Institut Neuromyogène, CNRS UMR 5310 – INSERM U1217, Université de Lyon, Lyon, France.

ACKNOWLEDGEMENT

The authors are indebted to late Pr Pierre Alain JOSEPH for having initiated the SOFMER Activity Score project and for his passionate guidance and constant support through the further steps of construction, scoring, and validation of this Score.

DISCLOSURE OF INTEREST

The authors declare that they have no competing interests.

Corresponding author

Dr Lorraine Charvolin

Service central de rééducation,

Hôpital Femme–Mère–Enfant, L'Escale

F-69500, Bron, France.

E-mail: Lorraine.charvolin@gmail.com

1 **Determining the inter-rater reliability of the SOFMER Activity Score (version 2) for**
2 **subjects in rehabilitation centers**

3 *Inter-rater reliability of SOFMER Score*

4

5 Abstract word count: 243

6 Text word count: 3909

7 Number of tables and figures: 5

8 Number of references: 28

9

10 **ABSTRACT**

11 **Objectives:** To assess the inter-rater reliability of the SOFMER Activity Score (SAS, version
12 2, an 8-item –4 motor and 4 cognitive– and 5-level scale) and improve its scoring system
13 before conducting further validation steps.

14 **Design:** Cross-sectional, prospective, observational, non-interventional, and multicentric
15 study.

16 **Setting:** The study was conducted between November 2018 and September 2019 in four
17 French rehabilitation centers (two public university hospitals for adults and two private not-
18 for-profit rehabilitation centers for children).

19 **Participants:** The study included 101 subjects (mean age: 44.5 years; SD: 25.4; 28.7% under
20 18 and 18.8% over 65). The female/male sex ratio was 0.6. The causes for admission to the
21 center were mainly neurological (65%) or orthopedic (24%).

22 **Interventions:** None.

23 **Main outcome measure:** Activity limitation was rated with the SOFMER Activity Score the
24 same day by two independent multidisciplinary teams. The inter-rater reliabilities of the Score
25 items were assessed using weighted kappa coefficients.

26 **Results:** All weighted kappa coefficients ranged between 0.83 and 0.92 indicating ‘good’ to
27 ‘excellent’ inter-rater reliability. Inter-team score disagreements occurred in 227 scores out of
28 808 (28%). The reason for most disagreements was unnoticed human or material aid during
29 the observation period.

30 **Conclusion:** The results demonstrate the high inter-rater reliability of the SASv2 and allow
31 carrying out further validation steps after minor changes to item scoring instructions and
32 clearer definitions of some items that help improving scoring standardization. The SASv2
33 may then become a consistent measure of activity level for clinical research or burden of care
34 investigations.

35

36 **KEYWORDS**

37 Activities of daily living, Rehabilitation, Rehabilitation centers, Reproducibility of results,
38 inter-rater reliability, SOFEMER Activity Score, Chronic limitation of activity

39

40 **ABBREVIATIONS**

41 FIM: Functional Independence Measure

42 HCP: health care provider

43 ICF: International Classification of Functioning, Disability, and Health

44 IRR: inter-rater reliability

45 RCs: Rehabilitation Centers

46 SAS: SOFEMER Activity Score

47 SASv2: SAS version 2

48

49 **INTRODUCTION**

50 In May 2001, the International Classification of Functioning, Disability, and Health (ICF)
51 “was officially endorsed by all 191 WHO Member States ... as the international standard to
52 describe and measure health and disability” [1]. In the ICF, disability (the antithesis of
53 functioning) refers to impairments of body structures and functions, limitations of activities,
54 and restrictions in participation (activities being tasks or actions an individual performs and
55 participation being the involvement in life situations). Functioning is further qualified by
56 distinguishing between capacity (what persons can do in a standard environment –test
57 conditions) and performance (what persons actually do in their usual environment –
58 community, home).

59 The ICF approach to disability integrated medical and social aspects into a ‘bio-
60 psycho-social’ model (including personal and environmental factors) and used new terms to
61 describe disability such as ‘impairments’, ‘limitations in activities’, and ‘restrictions in
62 participation’ [2]. The ICF identifies the necessary components of functioning, but does not
63 provide a measure to quantify functioning.

64 In rehabilitation medicine, frequent assessments of activity level in subjects with
65 disabilities are essential to anticipate activity loss, support personalized life projects, and
66 make clinical and management decisions. Among the current scales that assess subjects’
67 activities, some are activity-specific (the Functional Ambulation Category [3] or gait speed
68 tests that evaluate the gait but do not reflect overall activity levels) or population-specific (the
69 modified Rankin Scale for post-stroke neurological assessment [4] or the Instrumental
70 Activities of Daily Living adapted to geriatric patients [5]). Another scale is Barthel Index
71 used for “measuring changes in physical function of geriatric rehabilitation patients” [6, 7, 8]
72 or assessing functional recovery after hip fracture [9] or a neurological disorder such as stroke
73 [10]. In contrast, the Functional Independence Measure (FIM) is a general-purpose scale with

74 excellent psychometric properties [11] but is difficult to use in routine hospital care because
75 its administration requires 30 to 45 minutes [12].

76 Given these difficulties, the SOFMER Activity Score (SAS) was adapted from the ICF
77 in 2015 to assess accurately and rapidly the activity levels of subjects admitted to RCs
78 whatever their ages or clinical conditions. The SAS assigns independent scores to selected
79 motor and cognitive aspects of a subject's activity. Thus, though based on the concepts of the
80 ICF, the SAS is shorter and focuses on activity limitation in standardized environments,
81 whereas the ICF describes participation restriction in various individual and environmental
82 conditions.

83 An assessment scale has to undergo several tests to determine its strengths,
84 weaknesses, validity, responsiveness, and reliability [13]. The content validity of the SAS
85 version 1 (i.e., the relevance of its items to essential domains in medical rehabilitation) was
86 already established through three rounds of Delphi method [14] and its feasibility
87 demonstrated in a pilot study that involved 81 subjects. The latter assessment led to SAS
88 version 2 (SASv2) [14]. The validation process is ongoing.

89 In the process of a scale validation, 'reliability' is the reproducibility of the scale's
90 result over successive assessments, assuming the subject's condition has remained constant.
91 Reliability may take two forms: i) test-retest reliability, the reproducibility obtained by the
92 same investigator; and, ii) inter-rater reliability (IRR), the reproducibility obtained by
93 independent investigators assessing the same subject within a short period of time. The latter
94 form is essential because a high IRR is required for a confident use of the scale by various
95 health care providers (HCPs).

96 The aim of this study was to assess the IRR of the SASv2 in adult and pediatric
97 subjects in RCs and improve its scoring system, if necessary, before conducting further
98 validation steps.

99 **METHODS**

100 *The SOFMER Activity Score, version 2*

101 The SAS includes two domains: a Motor domain with items ‘Hygiene and dressing’,
102 ‘Feeding’, ‘Mobility’, and ‘Elimination’; and a Cognitive domain with items
103 ‘Communication’, ‘Relationships with others’, ‘Memory and knowledge translation’, and
104 ‘Task Execution’. Each item may be scored between 1 (the lowest score) and 5. A score of 1
105 represents ‘Activity impossible regardless of help’, a 2 ‘Activity possible with continuous
106 human help’, a 3 ‘Activity possible with human help or supervision’, a 4 ‘Activity possible
107 with technical help and/or adjustment but without human help’, and a 5 ‘Activity possible
108 without help’. The SAS provides instructions with examples to clarify the scoring process.

109

110 *Study design, setting, and participants*

111 The study was cross-sectional, prospective, observational, non-interventional, and
112 multicentric. Its objective was to determine the IRR of the SASv2 using ‘**Observer reported**
113 **outcome**’ (**ObsRO**) assessments [15].

114 The recruitment took place between November 2018 and September 2019. To be
115 eligible, all subjects had to be aged two years or more, to have been hospitalized for more
116 than four days in any of four French RCs (two public university hospitals for adults and two
117 private not-for-profit RCs for children), and to be able to give informed consent (personally or
118 via authorized persons). There were no exclusion criteria beyond those mentioned above.

119 All participants were solicited and enrolled by a physician during a stay at RC. They
120 were orally informed about the aim and the process of the study and hand-delivered an
121 information booklet. After consent, the physician collected the following data: age, sex, date,
122 and reason for RC admission.

123

124 *Study conduct and data collection*

125 To assess the IRR of the SASv2, each subject was scored on all eight items, the same day, by
126 two independent rater teams. The raters had to be HCPs from distinct professions (physician,
127 registered nurse, assistant nurse, therapist, etc.). The number of raters per subject and per team
128 had to be 2, 3, or 4 according to the availability of suitable raters.

129 The scoring process included no tests and no interviews; it was solely based on the
130 observation of subjects' abilities to perform everyday activities. The eight scores were
131 assigned according to what each subject was seen able to achieve during an at least four-day
132 stay in the RC. A single scoring form was filled out by each rating team; this required, on
133 average, 4.5 minutes per patient and was carried out as a team report during multidisciplinary
134 rounds. The dates and SASv2 scores were recorded together with the professions of the raters.
135 The rater teams were instructed not to communicate with each other until completion of data
136 collection.

137 The IRR analysis considered thus two series of SASv2 scores (one score per item, per
138 subject, and per rating team) and assessed the reliability between the two rater teams (not
139 between raters of same team).

140

141 *Statistical analyses*

142 According to the COSMIN Risk of Bias Checklist [16], a 'very good' assessment of the
143 SASv2 reliability requires a sample size greater than 100 subjects.

144 Each SASv2 item being ordinal with five levels, the IRR of each item was estimated
145 using a weighted kappa coefficient (κ_w) with its 95% confidence interval. This allows
146 expressing reliability as a number between 0 and 1 (0: no reliability; 1: perfect reliability).
147 Fleiss-Cohen weighting scheme (quadratic weights) was used to weight the disagreements
148 [17]. The results were interpreted as suggested by Landis & Koch [18]. Thus, $\kappa_w \geq 0.81$

149 indicated almost perfect agreement, $0.61 \leq \kappa_w \leq 0.80$ substantial agreement, $0.41 \leq \kappa_w \leq 0.60$
150 moderate agreement, $0.21 \leq \kappa_w \leq 0.40$ fair agreement, and $\kappa_w \leq 0.20$ slight agreement.

151 The observed frequencies and percentages of agreements or disagreements between
152 rater teams were examined once, for all eight items, in a single session. Exact and partial
153 agreements on each item were displayed on a Bangdiwala Chart [19] (Figure 1). This chart is
154 a representation that displays concordance in paired categorical data where areas of various
155 color densities represent exact and partial agreements. The Bangdiwala chart reflects also a
156 'joint distribution of the scores'; i.e., it gives a visual idea about the relative distributions of
157 the scores between the two rating teams.

158 The analysis examined also the distributions of the scores and floor or ceiling effects.
159 The latter terms are used when the scores are at or near the lower or upper limit, respectively
160 [20]. Herein, floor and ceiling effects relate to inflations of score 1 and score 5, respectively.

161 All statistical analyses were carried out with Statistical Analysis System software,
162 version 9.4. All tests were two-tailed and $p < 0.05$ was considered for statistical significance.

163

164 *Ethical considerations*

165 In accordance with the applicable regulations at the time of the study, a purely observational
166 study that did not change the management of the subjects/patients or required their active
167 participation needed neither a formal signed informed consent nor the agreement of an ethical
168 committee. Nevertheless, i) the investigators obtained verbal consents to the collection,
169 analysis, and publication of the study data; and, ii) the study received a favorable opinion
170 from the relevant ethics committee (Comité de Protection des Personnes Sud-Ouest et Outre-
171 Mer IV) on August 31, 2017. According to the current European guidelines (EU General Data
172 Protection Regulation), subjects' data for this research project were anonymized before

173 analysis and all data that could lead to participants' identification were kept confidential and
174 securely stored.

175 **RESULTS**

176 *Participants and raters*

177 Among 109 subjects originally included in all four RCs over eleven months, eight had to be
178 excluded because of one non-compliant rater team. No subjects were excluded after being
179 initially included. Thus, the study kept for analysis data on 101 subjects in whom the eight
180 SASv2 items were scored once by each rating team. As there were no missing scores, 808
181 data points were provided by each team and 808 pairs of scores could be compared.

182 The characteristics of the participants are displayed in Table 1. The mean (\pm SD) age
183 was 44.5 (\pm 25.4) years; 28.7% of the participants were under 18 and 18.8% over 70. The
184 female/male sex ratio was 60.3%. The participants were mainly admitted to RC for
185 neurological or orthopedic reasons (60.4 and 25.7%, respectively).

186 The raters of each team were for the most part nurses or assistant nurses. The
187 profession and number of the other raters depended on their availability at the time of scoring.
188 More precisely, the number of raters was 240 in Rating team 1 and 228 in Rating team 2. The
189 HCP occupations (numbers) in Rating teams 1 and 2 were respectively: nurses (105 and 100),
190 assistant nurses (93 and 89), pediatric nurses (21 and 21), rehabilitation physicians (10 and
191 10), physiotherapist (10 and 1), and medical students (1 and 7).

192

193 *Distribution of subjects' levels of activity on the SASv2*

194 Figure 1 shows that the distribution of the levels of activity varied widely across items. Level
195 5 was the most frequent except for items 'Hygiene and dressing' and 'Mobility'. Level 4 was
196 the least frequent especially for items 'Hygiene and dressing' and 'Elimination'. Level 1 was
197 poorly used for items 'Communication' and 'Relationships with others' and Level 2 poorly
198 used for item 'Memory'.

199 The score distributions varied widely by age group. The ceiling effect was less
200 important in subjects under 18 than in other age groups. Some levels were not represented in
201 the 19 subjects aged >70. Nearly all scores (Level 1 to Level 5) were assigned to each item.
202 No clear floor or ceiling effects were found; only domains ‘Feeding’, ‘Communication’, and
203 ‘Memory’ showed trends toward a ceiling effect (See Figures S1 and S2 in Supplementary
204 Material).

205

206 *Inter-rater reliability*

207 The percentage of disagreements between the two rating teams was 32.7% for ‘Hygiene and
208 dressing’, 23.8% for ‘Feeding’, 39.6% for ‘Mobility’, 27.7% for ‘Elimination’, 19.8% for
209 ‘Communication’, 27.7% for ‘Relationships with others’, 17.8% for ‘Memory’, and 34.7% for
210 ‘Task execution’.

211 The weighted kappa coefficients ranged from 0.83 to 0.92 (Figure 2). The lower
212 values concerned items ‘Relationships with others’ and ‘Task execution’ of the cognitive
213 domain ($\kappa_w = 0.83$) and item ‘Mobility’ of the motor domain ($\kappa_w = 0.84$). The less accurate
214 estimations (i.e., widest 95% CIs) concerned items ‘Task execution’, ‘Communication’, and
215 ‘Relationships with others’ (0.13, 0.15, 0.16, respectively, vs. 0.09 to 0.12 for the other
216 items).

217 Three out of four score disagreements (76.5%) were one-point differences (Tables 2
218 and 3). Of the 55 disagreements by more than one point, none reached a 4-point difference,
219 only 1 reached a 3-point difference (disagreement between Level 2 and 5). All others were 2-
220 point differences of which 65% were between Levels 3 and 5 (mainly concerning ‘Task
221 execution’ and ‘Relationships with others’), and 22% between Levels 1 and 3.

222 Score disagreements were the most frequent between Levels 2 and 3 for the motor
223 domain (mainly concerning ‘Hygiene and dressing’) and Levels 3 or 4 and 5 for the cognitive

224 domain (mainly concerning ‘Task execution’ and ‘Relationships with others’) (Tables 2 and
225 3).

226

227 *Disagreements and consensus scores*

228 After κ_w calculations, the rating teams compared their scores to determine the origins of any
229 disagreements and try to assign consensus scores.

230 On the 808 pairs of rates, there were 227 (28.1%) disagreements. No reason for
231 disagreement was found for 44 discordant score pairs (44/227; 19.4%), whereas a consensus
232 score could be assigned in 183 discordant score pairs (183/227; 80.6%).

233 In assigning the consensus scores, the lowest of the two scores was retained from 117
234 score pairs (117/183; 64%), the highest from 55 pairs (55/183; 30%), and an intermediate
235 whole number score in the remaining 11 pairs (11/183; 6%).

236 **DISCUSSION**

237 The present study reports on the IRR of the SOFMER Activity Score (SAS), a scale that
238 determines the activity level of subjects during medical rehabilitation in RCs. The IRR of any
239 measure of such status is important to ensure data consistency, which allows dependable
240 results and direct comparisons. Here, the weighted Kappa coefficients of agreement used to
241 compare two series of measurements made by two distinct rating teams in subjects with
242 various physical and/or mental impairments were “good” to “excellent”, ranging between κ_w
243 0.83 and 0.92.

244 As in the pilot study on the SAS [14], nearly all scores (Level 1 to Level 5) were used
245 for each item. Nevertheless, Level 4 was more frequently used than in the pilot study,
246 especially in the cognitive domain. Also, the absence of floor effect is important because it
247 allows assessing activity level improvements over time in the most severely impaired
248 subjects.

249 A future concurrent validity study is needed to determine whether the current SASv2
250 levels distinguish activity levels as well as the FIM, which is considered by some to be the
251 ‘gold standard’ for measuring function [21, 22] and is the most frequently used in French and
252 Swiss RCs. The FIM and the SASv2 were both developed from the ICF [1] (actually, the FIM
253 was developed from the old ICIDH –International Classification of Impairments, Disabilities,
254 and Handicaps). The current results confirm that the SASv2 is as reliable as the FIM [23].
255 This is supported by ‘almost perfect agreements’ [18] in item score comparisons between the
256 rating teams; all Kappa coefficients ranged from 0.83 to 0.92. According to Fleiss and Cohen
257 [17], when the scores are ordinal, Kappa coefficients can be interpreted as Intraclass
258 Correlation Coefficients (ICCs); thus, the SASv2 ‘Memory’ domain has a ‘very good
259 reliability’ ($\kappa_w \geq 0.91$), while the other domains have ‘good reliability’ (κ_w : 0.71 to 0.90) (0.71
260 and 0.90 are the ICC boundaries set by Fleiss and Cohen).

261 One explanation for the very high kappa values is that 76.5% of the disagreements
262 differed by one point only (the κ_w coefficient being weighted by the magnitude of the
263 disagreement between the raters). Another explanation is the effort made to standardize the
264 scoring with accurate definitions of the items and careful instructions on the scale use. For
265 instance, the scale requires a clear distinction between subject's performance and capacity;
266 whereas performance refers to the way a subject copes with disability in real-life situations
267 [24], capacity refers to the level of activity a subject may reach in a standard environment
268 without assistance and represents the HCP's idea of the goals to reach. One advantage of the
269 ICF over the SASv2 is that it explores both concepts; still, the SAS was created to focus on
270 the daily performance of the subjects.

271 For standardization purposes, the SASv2 instructions underline that a rating team
272 should include HCPs from different professions to ensure a variety of opinions and scores
273 regarding activity limitation [25]. Furthermore, the instructions insist on a four-day
274 observation period. This four-day period has been initially set as: i) the minimum residence
275 time in a RC for subjects inclusion; ii) the time sufficient to allow subject observation in
276 various circumstances by at least two different HCPs; and, iii) the standard time for the
277 successive scale validation steps. This relatively short observation period contributed
278 probably to the high IRRs. It was important that the subjects did not change over the study
279 period, as this would have compromised the testing reliability. Actually, in a previous study
280 [18], observations over longer periods –during which slight or moderate changes in the
281 subjects' clinical conditions occurred– have resulted in less satisfactory IRRs.

282 The study showed that, in the motor domain, most disagreements concerned Levels 2
283 and 3 although these levels were not over-represented. The raters related the disagreements to
284 some lack of clarity about the meaning of 'supervision' in the definition of Level 3. Indeed,
285 'supervision' would suggest the need for assistance with all or part of a given activity. We

286 suggest thus clarifying the meaning of 'supervision' in the definitions of Levels 2 and 3
287 ('Activity possible with continuous human help or supervision' and 'Activity possible with
288 partial human help or supervision'). In the cognitive domain, most disagreements concerned
289 Levels 4 and 5; this might be due to an as yet unexplained over-representation of Level 5. In
290 many cases, the raters' explanation was the lack of human assistance (e.g., subject 'unable to
291 cope with night needs', 'seeks help', 'needs to be stimulated'). This and the fact that the
292 consensus on the final scores were set to lower scores in 65% of all disagreements indicate
293 that a high proportion of scores failed to take into account the subject's whole environment
294 (e.g., use of wheelchair, sit-to-stand lift, or braces or need for human help in transfers or
295 diaper use).

296 The discussions during the consensus meetings led to better SASv2 standardization.
297 This meant: i) more accurate definitions of 'Hygiene and dressing' that excludes now the
298 notion of transfers; ii) clearer examples of SASv2 items that allow for the use of new objects
299 or aspects; e.g., equipment for 'Elimination', withdrawal for 'Relationships with others', and
300 acting according to one's will for 'Task execution'; iii) additional and more accurate
301 examples, especially regarding 'Memory' and 'Relationships with others'; and, iv) a
302 suggestion for using a clearer scoring system (See the online Appendix). Indeed, standardized
303 scales have the advantages of controlling for the variety of impairments and disabilities that
304 affect functional assessment, reducing scoring errors, and ensuring effective and consistent
305 scale use various institutions.

306 As stated above, Kappa coefficients can be interpreted as ICCs [17, 26]. Here, we
307 compare ICCs between various scales, even though IRRs should be compared only between
308 scales with similar aims, domains, items, etc. The mean IRRs of the SASv2 items (all >0.82)
309 compare well with those of the FIM items that ranged between 0.57 and 0.85 with only three
310 of those 18 items having IRRs >0.80 [23]. In addition, a review about Barthel index reported

311 excellent IRR (0.93) in stroke patients [10] but only low-to-moderate IRRs in the elderly and
312 even worst results in subjects with cognitive impairment [27]. A different review reported that
313 the IRRs of the modified Barthel index ranged between 0.25 and 0.95 [28]. Thus, despite
314 various differences, the SASv2 compares favorably with other known scales. However, as
315 cognitive impairments can decrease functional abilities, it would be interesting to compare
316 motor domain scores between SASv2, Barthel index, and FIM in cognitively intact vs.
317 cognitively impaired subjects.

318

319 *Assets*

320 One asset of the SASv2 is its immediate, accurate, and reliable use by HCPs. Indeed, using
321 the SASv2 does not require formal training because the scale instructions for use were
322 initially specifically designed and deemed sufficiently clear to be satisfactorily implemented
323 by any HCP. This was proven by the good inter-rater reproducibility seen here. Nevertheless,
324 the successive validation steps may suggest introducing minor amendments for even better
325 implementation.

326 Additionally, the SASv2 has proven to be less time consuming than other scales. In
327 fact, the raters do not have to scrutinize every aspect of every subject as in other measures in
328 which timed, planned, and targeted observations are required. They do not have to dedicate
329 professional time to those kinds of observations; they just have to state their scores on a
330 subject's activity level after a passive observation of more than four days.

331 Finally, the FIM has two versions, the FIM (for adults) and the WeeFIM (for children
332 aged six months to seven years), whereas the SASv2 covers children without the need for a
333 separate measure.

334

335

336 *Limitations*

337 In this study, the number of raters per team (2 to 4) was significantly lower than in the pilot
338 study (mean: 6.4, range: 2-11) [14]. The explanation is that each RC had to recruit two rating
339 teams; this i) decreased the number of potential raters per team; ii) reduced the benefits from
340 larger teams in terms of observation accuracy; and, iii) increased the risks of errors and
341 omissions. Obviously, the higher the number of raters, the higher the IRR. Thus, we
342 recommend each rating team include at least three HCPs (see online Appendix).

343 The predominance of nurses and assistant nurses as raters helped obtaining good
344 agreements between raters. Nevertheless, this reflects the reality of the subjects' environment;
345 these HCPs are those who are in frequent daily contact with several subjects within a given
346 RC. The other professionals i) may not have to be in (sufficient) contact with some categories
347 of subjects during their stay (short or irregular care sessions); and, ii) may not be as available
348 as nurses or assistant nurses. This implies seeking, as far as possible, the participation of
349 raters other than nurses.

350 The IRR is an important early step in the process of scale development. Whether that
351 reliability may differ with subjects' diseases or other factors is certainly an interesting issue
352 but requires other study designs. Another fact of the SASv2 to be considered is the ideal of
353 domain subscores or total score. In principle, the contents of the domains are so varied that a
354 total score might not be relevant in terms of activity level. Nevertheless, potential uses of
355 those scores will be the topics of future studies. Additionally, an analysis may determine
356 whether there is a correlation between the SASv2 total score and the burden of care.

357 At present, the high IRR of the SASv2 (or its consistency) in RC residents allows
358 evaluating and comparing subjects' activity levels. In the future, it will allow setting health-
359 status improvement objectives, improving management, anticipating activity limitation, and

360 planning hospital discharge. In addition, accurate measurements of activity levels may reflect
361 the burden of care and help hospital managers improve staffing.

362

363 *Conclusions*

364 This study succeeded in assessing the IRR of the SASv2 in adult and pediatric subjects
365 admitted to RCs. All IRRs were 0.83 or higher, which indicated 'good' reliability.

366 Discussions on score disagreements improved slightly the previous version of the scale.

367 In next steps, other important psychometric properties of the SASv2 have to be
368 investigated in multicenter studies: construct validity, criterion validity, convergent validity,
369 test-retest reliability, and responsiveness (or sensitivity to change). These validation steps will
370 provide strong arguments in favor of replacement of other scales that would prove less valid
371 or more time-consuming.

372 With hopefully successive encouraging results, the SASv2 will prove useful not only
373 for improving and planning care but also for designing clinical trials because the ability to
374 form homogeneous groups of subjects using the SASv2 (or another scale) is essential for
375 testing the efficacy of new drugs or interventions.

376 **REFERENCES**

- 377 [1] WHO. International Classification of Functioning, Disability and Health (ICF).
378 <http://www.who.int/classifications/icf/en>.
- 379 [2] Stucki G, Cieza A, Melvin J. The International Classification of Functioning, Disability
380 and Health (ICF): a unifying model for the conceptual description of the rehabilitation
381 strategy. *J Rehabil Med* 2007;39:279-85. <https://doi.org/10.2340/16501977-0041>.
- 382 [3] Holden MK, Gill KM, Magliozzi MR, Nathan J, Piehl-Baker L. Clinical gait assessment
383 in the neurologically impaired. Reliability and meaningfulness. *Phys Ther* 1984;64:35–
384 40. <https://doi.org/10.1093/ptj/64.1.35>.
- 385 [4] Wilson JTL, Hareendran A, Grant M, Baird T, Schulz UGR, Muir KW, et al. Improving
386 the assessment of outcomes in stroke: use of a structured interview to assign grades on
387 the modified Rankin Scale. *Stroke* 2002;33:2243–6.
388 <https://doi.org/10.1161/01.str.0000027437.22450.bd>
- 389 [5] Lawton MP, Brody EM. Assessment of older people: self-maintaining and instrumental
390 activities of daily living. *Gerontologist* 1969;9:179-86.
391 http://doi.org/10.1093/geront/9.3_Part_1.179
- 392 [6] Lam SC, Lee DTF, Yu DSF. Establishing CUTOFF values for the Simplified Barthel
393 Index in elderly adults in residential care homes. *J Am Geriatr Soc* 2014;62:575–7.
394 <https://doi.org/10.1111/jgs.12716>.
- 395 [7] Bouwstra H, Smit EB, Wattel EM, van der Wouden JC, Hertogh CPM, Terluin B, et
396 al. Measurement Properties of the Barthel Index in Geriatric Rehabilitation. *J Am Med*
397 *Dir Assoc* 2019;20:420-425.e1. <https://doi.org/10.1016/j.jamda.2018.09.033>.
- 398 [8] Smit EB, Bouwstra H, van der Wouden JC, Hertogh CPM, Wattel EM, Roorda LD,
399 Terwee CB. Development of a Patient-Reported Outcomes Measurement Information
400 System (PROMIS®) short form for measuring physical function in geriatric

401 rehabilitation patients. *Qual Life Res* 2020;29:2563-72. <https://doi.org/10.1007/s11136->
402 020-02506-5

403 [9] Mayoral AP, Ibarz E, Gracia L, Mateo J, Herrera A. The use of Barthel index for the
404 assessment of the functional recovery after osteoporotic hip fracture: One year follow-
405 up. *PLoS One* 2019;14:e0212000. <https://doi.org/10.1371/journal.pone.0212000>

406 [10] Duffy L, Gajree S, Langhorne P, Stott DJ, Quinn TJ. Reliability (inter-rater agreement)
407 of the Barthel Index for assessment of stroke survivors: systematic review and meta-
408 analysis. *Stroke* 2013;44:462–8. <https://doi.org/10.1161/STROKEAHA.112.678615>.

409 [11] Dodds TA, Martin DP, Stolov WC, Deyo RA. A validation of the functional
410 independence measurement and its performance among rehabilitation inpatients. *Arch*
411 *Phys Med Rehabil* 1993;74:531-6. [https://doi.org/10.1016/0003-9993\(93\)90119-u](https://doi.org/10.1016/0003-9993(93)90119-u).

412 [12] Granger CV, Hamilton BB, Keith RA, Zielezny M, Sherwin FS. Advances in functional
413 assessment for medical rehabilitation. *Top Geriatr Rehabil* 1986;1:59-74. No doi found.

414 [13] Health measurement scales: a practical guide to their development and use (5th edition).
415 *Aust N Z J Public Health* 2016;40:294-5. <https://doi.org/10.1111/1753-6405.12484>.

416 [14] Morard MD, Gonzalez-Monge S, Rippert P, Roche S, Bernard JC, Lagauche D, et al.
417 Construction and feasibility study of the SOFMER Activity Score (SAS), a new
418 assessment of physical and cognitive activity. *Ann Phys Rehabil Med* 2018;61:315-22.
419 <https://doi.org/10.1016/j.rehab.2018.04.006>.

420 [15] Walton MK, Powers JH, Hobart J, Patrick DL, Marquis P, Vamvakas S, et al. Clinical
421 Outcome Assessments: Conceptual Foundation–Report of the ISPOR Clinical Outcomes
422 Assessment – Emerging Good Practices for Outcomes Research Task Force DOES THIS
423 HAVE TO BE LABELED AS PART 1. *Value Health*. 2015;18:741-52.
424 <https://doi.org/10.1016/j.jval.2015.08.006>.

- 425 [16] COSMIN Risk of Bias Checklist. [https://www.cosmin.nl/wp-content/uploads/COSMIN-](https://www.cosmin.nl/wp-content/uploads/COSMIN-RoB-checklist-V2-0-v17_rev3.pdf)
426 [RoB-checklist-V2-0-v17_rev3.pdf](https://www.cosmin.nl/wp-content/uploads/COSMIN-RoB-checklist-V2-0-v17_rev3.pdf)
- 427 [17] Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation
428 coefficient as measures of reliability. *Educ Psychol Meas* 1973;33:613-9.
429 <https://doi.org/10.1177/001316447303300309>.
- 430 [18] Landis JR, Koch GG. The measurement of observer agreement for categorical data.
431 *Biometrics* 1977;33:159-74. <http://doi.org/10.2307/2529310>
- 432 [19] Bangdiwala SI, Shankar V. The agreement chart. *BMC Med Res Methodol* 2013;13:97.
433 <https://doi.org/10.1186/1471-2288-13-97>.
- 434 [20] Everitt BS. *The Cambridge dictionary of statistics* (2nd ed.). Cambridge, UK:
435 Cambridge University Press. 2002.
- 436 [21] Mount Sinai Rehabilitation Center. 2017 Scorecard Inpatient Rehabilitation.
437 [https://www.mountsinai.org/files/MSHealth/Assets/HS/Care/Rehab-Medicine/Inpatient-](https://www.mountsinai.org/files/MSHealth/Assets/HS/Care/Rehab-Medicine/Inpatient-Services/2017Overall-Rehabilitation-Center-Outcomes.pdf)
438 [Services/2017Overall-Rehabilitation-Center-Outcomes.pdf](https://www.mountsinai.org/files/MSHealth/Assets/HS/Care/Rehab-Medicine/Inpatient-Services/2017Overall-Rehabilitation-Center-Outcomes.pdf)
- 439 [22] Velozo CA, Byers KL, Wang YC, Joseph BR. Translating measures across the
440 continuum of care: Using Rasch analysis to create a crosswalk between the Functional
441 Independence Measure and the Minimum Data Set Craig A. *J Rehab Res Dev*
442 2007;44:467-78. <https://doi.org/10.1682/JRRD.2006.06.0068>
- 443 [23] Ottenbacher KJ, Hsu Y, Granger CV, Fiedler RC. The reliability of the functional
444 independence measure: a quantitative review. *Arch Phys Med Rehabil* 1996;77:1226-32.
445 [https://doi.org/10.1016/s0003-9993\(96\)90184-7](https://doi.org/10.1016/s0003-9993(96)90184-7).
- 446 [24] Tarvonen-Schröder S, Laimi K, Kauko T, Saltychev M. Concepts of Capacity and
447 Performance in Assessment of Functioning Amongst Stroke Survivors: A Comparison of
448 the Functional Independence Measure and the International Classification of

449 Functioning, Disability and Health. *J Rehabil Med* 2015;47.
450 <https://doi.org/10.2340/16501977-1974>.

451 [25] Saltychev M, Tarvonen-Schröder S, Bärlund E, Laimi K. Differences between
452 rehabilitation team, rehabilitants, and significant others in opinions on functioning of
453 subacute stroke survivors: Turku ICF study. *Int J Rehabil Res* 2014;37:229-35.
454 <https://doi.org/10.1097/MRR.0000000000000065>.

455 [26] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol*
456 *Bull* 1979;86:420-8. <https://doi.org/10.1037//0033-2909.86.2.420>.

457 [27] Sainsbury A, Seebass G, Bansal A, Young JB. Reliability of the Barthel Index when
458 used with older people. *Age Ageing* 2005;34:228–32.
459 <https://doi.org/10.1093/ageing/afi063>.

460 [28] Quinn TJ, Dawson J, Walters MR, Lees KR. Reliability of the modified Rankin Scale: a
461 systematic review. *Stroke* 2009;40:3393–5.
462 <https://doi.org/10.1161/STROKEAHA.109.557256>.

463

464 **FUNDING**

465 Funding was provided by the *Société française de médecine physique et de réadaptation*
466 (SOFMER) and the French Ministry of Health (Direction générale de l'offre de soins, DGOS)
467 within the context of *Programme de recherche sur la performance du système de soins* (grant
468 no. PREPS-16-390).

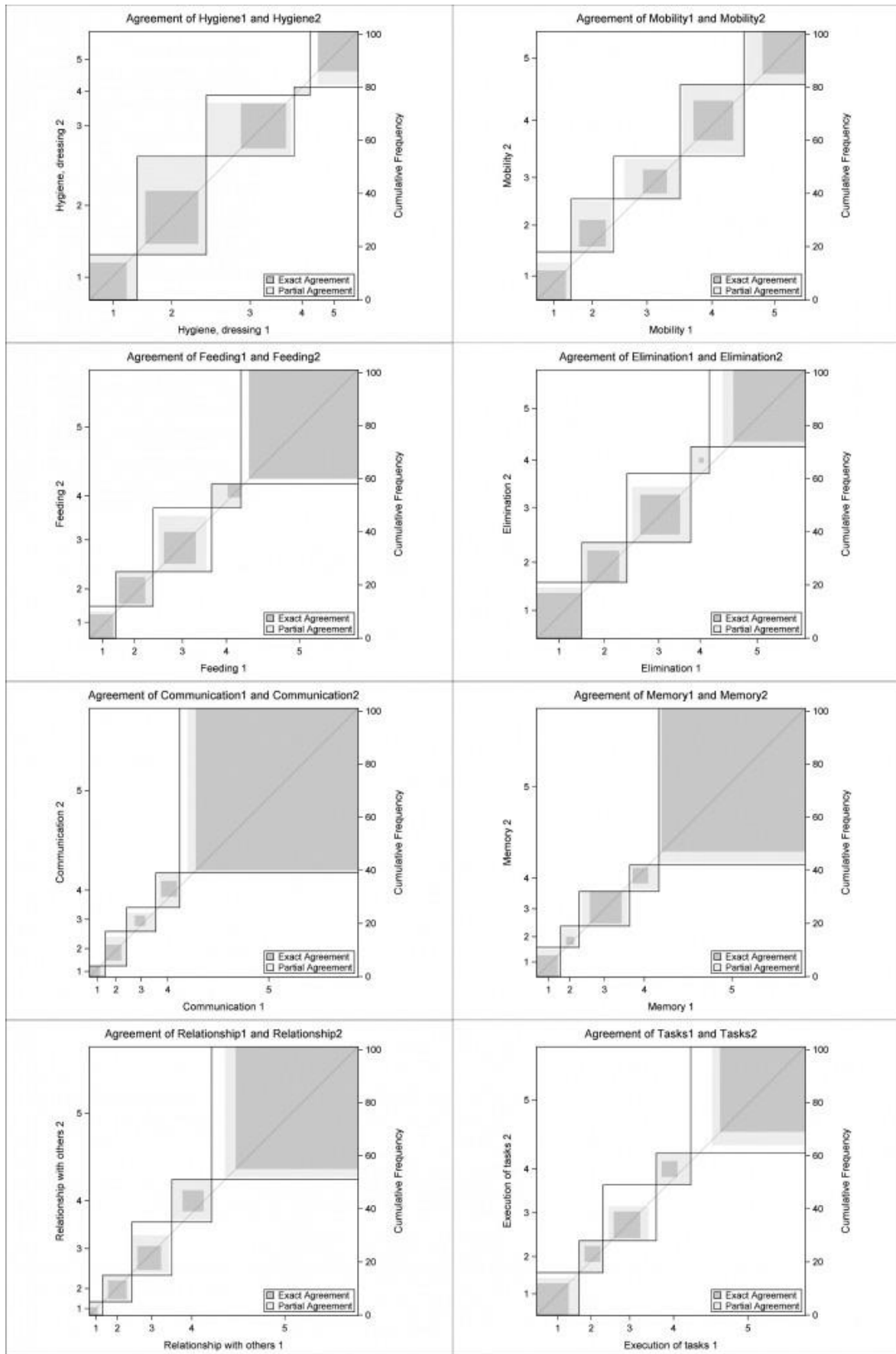
469 **LEGENDS TO THE FIGURES**

470

471 Figure 1 – Bangdiwala’s agreement chart comparing SASv2 scores between the two rating
472 teams. Each item is represented in a distinct panel in which the five levels are represented by
473 rectangles with one to three shades of grey. A deep grey area represents an exact agreement, a
474 light grey area a partial agreement with a ‘one level away’ discrepancy, and a white area a
475 partial agreement with a ‘two-level away or more’ discrepancy. Mentions “1” and “2” refer to
476 “Rating team 1” and “Rating team 2”.

477

478 Figure 2 – Weighted kappa coefficients of agreement with their 95% confidence intervals.



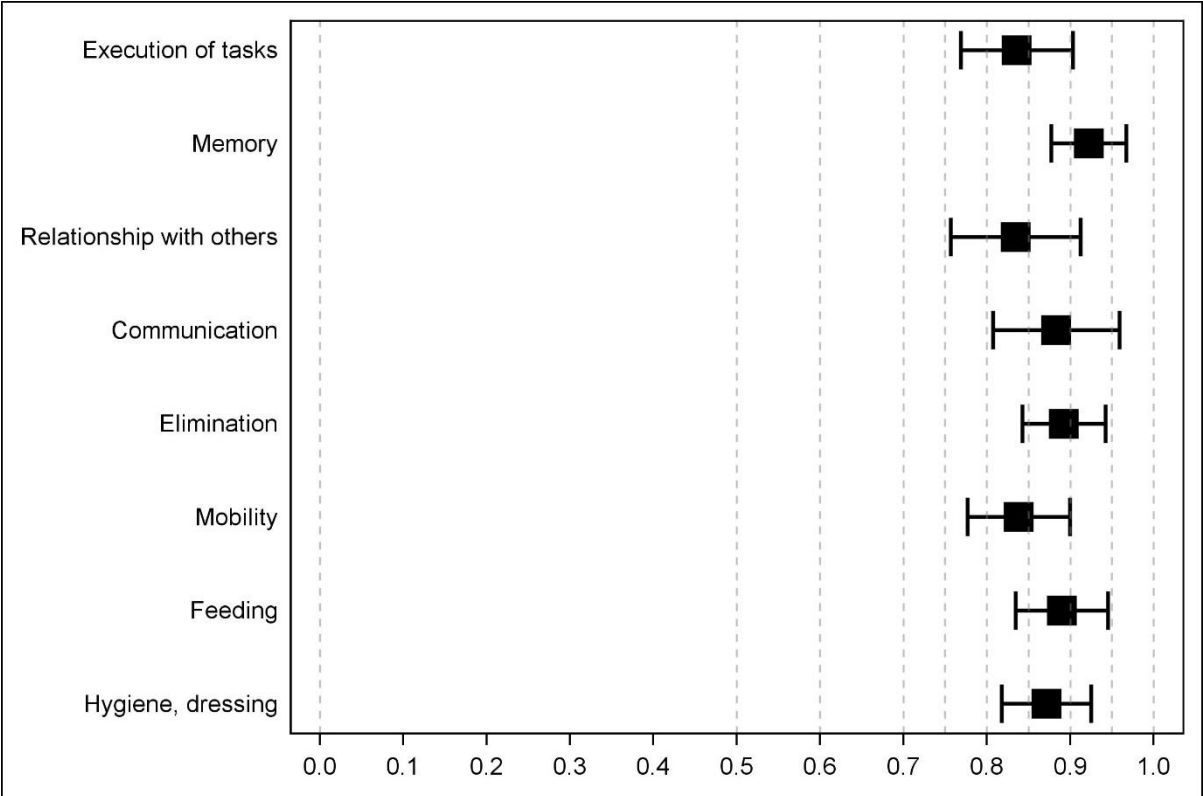


Table 1 - Characteristics of the participants in the study of inter-rater reliability of the SOFMER Activity Score (SAS) (n=101)

Characteristic	Number (percentage)
Age category	
<18 years	29 (28.71)
18 – 70 years	53 (52.48)
>70 years	19 (18.81)
Sex	
Females	38 (37.62)
Males	63 (62.38)
Reason for hospital admission	
Neurological of central origin	
Stroke	55 (54.45)
Stroke	23 (22.77)
Cerebral palsy	5 (4.95)
Spinal cord injury	8 (7.92)
Head trauma	6 (5.94)
Degenerative disease	1 (0.99)
Parkinson disease	1 (0.99)
Multiple sclerosis	2 (1.98)
Tumor and malformation	4 (3.96)
Other ¹	5 (4.95)
Neurological of peripheral origin	
6	6 (5.94)
Orthopedic	
26	26 (25.74)
Prosthesis	7 (6.93)
Fracture	6 (5.94)
Traumatic injuries	5 (4.95)
Agenesis	4 (3.96)
Other ²	4 (3.96)
Cardiopulmonary	2 (1.98)
Rheumatological	2 (1.98)
Bedsore	3 (2.98)
Other causes for hospitalization ³	5 (4.95)
Unspecified cause for hospitalization	2 (1.98)

¹ Unspecified hemiplegia, hemorrhage, or anoxic cerebral lesion - ² Osteochondrodysplasia, unspecified intervention, clubfoot, spondylolisthesis - ³ Dissociative amnesia, extreme immaturity, sphingolipidosis, congenital multiple exostoses, Marfan syndrome.

Table 2 – Number of exact and partial agreements regarding the items of the motor domain (101 patients).

Type of agreement	Motor domain items			
	Hygiene, dressing	Feeding	Mobility	Elimination
Exact agreements				
Level 1	14	9	11	17
Level 2	20	10	10	12
Level 3	17	12	9	15
Level 4	1	5	15	2
Level 5	15	41	16	27
Disagreements				
Level 1 vs. 2	7	2	5	2
Level 2 vs. 3	16	5	9	6
Level 3 vs. 4	2	10	9	7
Level 5 vs. 4	5	0	10	6
Level 2 vs. 4	0	0	2	0
Level 3 vs. 1	0	2	4	2
Level 5 vs. 3	4	5	1	5
Level 5 vs. 2	0	0	0	0

Table 3 – Number of exact and partial agreements regarding the items of the cognitive domain (101 patients).

Type of agreement	Cognitive domain items			
	Communication	Relationships	Memory	Task execution
Exact agreement				
Level 1	4	3	8	12
Level 2	6	7	3	6
Level 3	4	9	12	10
Level 4	6	8	6	6
Level 5	61	46	54	32
Disagreement				
Level 1 vs. 2	2	3	3	6
Level 2 vs. 3	5	3	4	3
Level 3 vs. 4	5	7	2	5
Level 5 vs. 4	4	7	5	8
Level 2 vs. 4	1	1	2	0
Level 3 vs. 1	0	1	1	2
Level 5 vs. 3	2	6	1	11
Level 5 vs. 2	1	0	0	0