


ORIGINAL WORK



Cross-Cultural Adaptation and Validation of the Greek Version of the “Full Outline of Unresponsiveness Score”: A Prospective Observational Clinimetric Study in Neurosurgical Patients

Dimitrios M. Anestis^{1*} , Parmenion P. Tsitsopoulos¹, Nikolaos G. Foroglou², Marianna S. Tsatali³, Konstantinos Marinos¹, Marios Theologou¹ and Christos A. Tsonidis¹

© 2021 Springer Science+Business Media, LLC, part of Springer Nature and Neurocritical Care Society

Abstract

Background: The Full Outline of Unresponsiveness (FOUR) score is a clinical instrument for the assessment of consciousness which is gradually gaining ground in clinical practice, as it incorporates more complete neurological information than the Glasgow Coma Scale (GCS). The main objectives of the current study were the following: (1) translate and cross-culturally adapt the FOUR score into Greek; (2) evaluate its clinimetric properties, including interrater reliability, internal consistency, and construct validity; and (3) evaluate the reliability of assessments among health care professionals with different levels of experience and training.

Methods: The FOUR score was initially translated into Greek. Next, patients with neurosurgical pathologies in need of consciousness monitoring were independently assessed with the GCS and FOUR score within one hour by four raters who had different levels of experience and training (two board-certified neurosurgeons, a neurosurgery resident, and a registered nurse). Interrater reliability, internal consistency, and construct validity were evaluated for the scales using weighted Cohen's κ (κ_w) and intraclass correlation coefficients (ICC), Cronbach's α and Spearman's ρ values, respectively.

Results: A total of 408 assessments were performed for 99 patients. The interrater reliability was excellent for both the FOUR score (ICC = 0.941) and GCS (ICC = 0.936). The values of κ_w exceeded 0.90 for all pairs, suggesting that the FOUR score can be reliably applied by raters with varying experience. Among the scales' components, FOUR score's brainstem and respiratory items showed the lowest, yet high enough ($\kappa_w > 0.60$), level of agreement. The interrater reliability remained excellent ($\kappa_w > 0.85$, ICC > 0.90) for all diagnosis and age groups, with a trend toward higher FOUR score values in the most severe cases (ICC = 0.813 vs. 0.723). Both the FOUR score and GCS showed high internal consistency (Cronbach's $\alpha > 0.70$ for all occasions). The FOUR score correlated strongly with GCS (Spearman's $\rho > 0.90$ for all raters), suggesting high construct validity.

*Correspondence: dimanestis@yahoo.gr

¹ Department of Neurosurgery, Hippokratia General Hospital, Aristotle University School of Medicine, 49 Konstantinoupoleos Str., 54642 Thessaloniki, Greece

Full list of author information is available at the end of the article

Conclusions: The Greek version of the FOUR score is a valid and reliable tool for the clinical assessment of patients with disorders of consciousness. It can be applied successfully by nurses, residents, and specialized physicians. Therefore, its use by medical practitioners with different levels of experience and training is strongly encouraged.

Keywords: Coma scale, Level of consciousness, Full outline of unresponsiveness score, Glasgow coma scale, Interrater reliability, Validation

Introduction

Assessment of the level of consciousness and the depth of coma are important aspects of neurological examination [1]. The Glasgow Coma Scale (GCS) is the clinical instrument most widely used for this purpose, assessing eye opening, verbal response, and motor response [2–6]. Nevertheless, numerous downsides have been outlined over the years, such as the inability to reliably evaluate intubated patients or patients with tracheostomy and the lack of evaluation of brainstem reflexes [1, 7–12]. To overcome some of those drawbacks, a number of scales, such as the Reaction Level Scale (RLS85) and the Innsbruck Coma Scale, have been proposed in clinical practice [1].

The Full Outline of Unresponsiveness (FOUR) score seems to be of particular interest, since it is increasingly implemented worldwide and already available in many languages, yet not in Greek [13]. This scale does not include a verbal component (an element that cannot be always reliably assessed), but it contains other important clinical indicators, such as the assessment of the patient's breathing pattern and brainstem reflexes [14].

The exclusion of verbal assessment has been presented as a potential advantage by the FOUR score's originators [14]. However, because this component cannot be assessed in patients in severe condition (especially those under mechanical ventilation), its significance may be limited for patients with mild or moderate disturbance of responsiveness, in which verbal performance is considered a critical distinguishing feature. In those patients, the GCS still has high clinical value [15, 16].

The original validation of the FOUR score found that agreement between nurses when applying the scale was "less than optimal" [14, 17]. Since then, the interrater reliability of the scale has been assessed by various reports with good results overall [18]. However, studies that involve a large number of neurological assessments are limited [19–21]. Moreover, in reports that included evaluators with different levels of experience and training, patients were not assessed consistently by all available raters.

This study aimed to: (1) translate and cross-culturally adapt the FOUR score into Greek; (2) evaluate the FOUR score's interrater reliability, internal consistency, and construct validity; and (3) test the hypothesis that the

FOUR score can be consistently applied reliably by raters with different levels of health care education and experience. The secondary objective of this study was the evaluation of the interrater reliability and internal consistency for the Greek version of the GCS, in comparison with the FOUR score.

Methods

Development of the Greek Version

The "Process of Translation and Adaptation of Instruments" proposed by the World Health Organization was followed for the development of the Greek version of the scale [22].

Initially, the scale was translated into Greek by author DMA, who is a native Greek speaker neurosurgeon, proficient in English, and well familiar with the required scientific terminology. Next, an expert panel was formed by authors PPT, NGE, and CAT, (neurosurgeons fluent in either language), a psychologist with experience in the development of Greek translations of clinical instruments from English (MST), and the original translator (DMA). In this step, all discrepancies regarding the clarity of the translation were resolved. The final version was translated back to English by a bilingual health care professional based in the United Kingdom, who was blinded to the original scale. The back-translated version was then compared with the original one by the expert panel and by Professor EFM Wijdicks, the originator of the FOUR score (Table 1).

The GCS is considered the gold standard clinical tool for the assessment of coma [4–6]; thus, it was decided to use the GCS for comparison with the FOUR score. The examiners used the existing Greek version of the GCS [23], which is the sole coma scale used in Greek hospitals. The GCS is being taught to medical students, nurse students, and residents throughout the country and has been included in a number of official medical textbooks (Internal Medicine, Intensive Care Medicine, Neurology, and Neurosurgery). Thus, its adoption by Greek medical practitioners has been vast for decades. To avoid unnecessary disruptions and inconveniences during the study of its clinimetric parameters and the comparison with those of the FOUR score, it was decided to keep the existing Greek version of the GCS.

Table 1 The Full Outline of Unresponsiveness score

Eye response	
4	Eyelids open or opened, tracking, or blinking to command
3	Eyelids open but not tracking
2	Eyelids closed but open to loud voice
1	Eyelids closed but open to pain
0	Eyelids remain closed with pain
Motor response	
4	Thumbs-up, fist, or peace sign
3	Localizing to pain
2	Flexion response to pain
1	Extension response to pain
0	No response to pain or generalized myoclonus status
Brainstem reflexes	
4	Pupil and corneal reflexes present
3	One pupil wide and fixed
2	Pupil or corneal reflexes absent
1	Pupil and corneal reflexes absent
0	Absent pupil, corneal, and cough reflex
Respiration	
4	Not intubated, regular breathing pattern
3	Not intubated, Cheyne–Stokes breathing pattern
2	Not intubated, irregular breathing
1	Breathes above ventilator rate
0	Breathes at ventilator rate or apnea

To ensure consistency, written guidelines for the application of the two scales (based on our previous experience and the existing literature) were also formed and distributed to the evaluators during their assessments.

Validation in Neurosurgical Patients

Study Design and Setting

A prospective observational cohort study was conducted. The protocol had been previously defined and approved by all authors. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) cohort reporting guidelines [24] and the Consensus-based Standards for the Selection of Health Measurement Instruments (COSMIN) taxonomy and methodology [25–28] were followed.

The study was conducted between October 1st, 2018, and December 31st, 2020, in the Department of Neurosurgery at Hippokraton General Hospital, Thessaloniki, Greece, which is a 24-bed unit, including a 20-bed general ward and a 4-bed critical care unit. The latter critical care unit involves patients in need of closer monitoring and more intense treatment, such as those with markedly disturbed level of consciousness, potential candidates for clinical deterioration, individuals coming from long

hospitalization in the intensive care unit, and those with a tracheostomy. Patients treated in the intensive care unit of the hospital were also included in our cohort, provided that they presented with neurosurgical pathology. The study protocol was approved by the ethics committee of the hospital (Ref. Nr. 985-2017). In compliance with the current legislation, the National Data Protection Authority was notified on its conduction (Ref. Nr. 850-2018). The study was conducted in accordance with the ethical standards of the 1964 Helsinki Declaration and its later amendments.

To reach the best possible level in terms of medical ethics, legal consent was obtained from patients capable of providing it or by proxy when deemed necessary.

Variables and Data Collection

As noted above, the level of consciousness was assessed using the GCS and the scale under validation (FOUR score). All participants were examined by four raters: two board-certified neurosurgeons, one senior (PPT) and one junior (DMA), a resident of neurosurgery (six in total) and a nurse (eight in total). All four assessments of each rating session were performed independently by the evaluators, and within one hour, ensuring that patients' neurological status did not change during this period. Patients in which the level of consciousness was altered during a rating session (as noted by their supervising physician) were assessed anew by all examiners. Evaluators were blinded to other raters' results.

The residents were equally distributed through all possible years of training in neurosurgery in our department (from the first to the last) and the involved nurses graduated from accredited nursing schools, with a clinical experience ranging from 10 to 25 years. None had any prior experience on the application of the FOUR score. All were previously informed in detail on the methodology and the objectives of the study and received vigorous training on the application of the scales by the two certified neurosurgeons. The training involved a one-hour lecture and ten exhibition assessments by the trainers in the initial phase, and, in the final phase, 30 assessments on acting patients, with predefined scenarios of varying severity, covering all possible subscores for each component of the two scales. For the FOUR score's brainstem subscores from 0 to 3 and respiratory 0 and 1, additional training was performed on actual patients. This procedure took place in the intensive care unit and on patients in critical condition in the Neurosurgery Department, after obtaining informed consent by a next-of-kin. It was a prerequisite for each rater to reach a consensus with the supervisors for a number of assessments (including subscores) before official patient recruitment. All raters were given in written form the two scales and application

instructions in Greek, which they were asked to use throughout their training and study conduction.

For each patient, the results from the assessments of the level of consciousness were collected directly. The remaining data of interest (sex, age, diagnosis, clinical course, and imaging findings) were extracted during hospitalization or directly after each participant's discharge. All data were completely anonymized and digitally recorded in a Microsoft Excel 2019 (Microsoft Corporation, Redmond, WA) spreadsheet. The procedure was in full compliance with the current legislation.

Eligibility Criteria

Patients (1) 18 years old or older, (2) treated in the Department of Neurosurgery, and (3) in need of consciousness monitoring were enrolled in the study. This included participants presenting on admission with disturbed, but also with normal level of consciousness that could be disrupted during their hospitalization, to avoid bias during the assessment of alert patients.

Exclusion criteria were the following: (1) absence of legal consent (denial, withdrawal or inability to acquire), (2) unavailability of trained examiners to obtain complete patient assessment within one hour, (3) clinical deterioration during one rating session (between the assessment of different examiners) in case of inability to acquire anew all ratings reliably, (4) patients with dementia or mental illness, (5) patients under sedatives, neuromuscular junction blockers, alcohol or addictive substances, that could not be reliably evaluated, and (6) cases with missing data. Hospitalization in other departments, including the intensive care unit, was not a cause of exclusion.

Statistical Analysis

Descriptive statistics are presented as means \pm standard deviation or medians. Normality of data was checked with the Kolmogorov–Smirnov test.

Quadratic weighted Cohen's kappa (κ_w) values and intraclass correlation coefficients (two-way random-effects, single rater, consistency ICC) with a confidence interval of 95% were calculated to assess interrater reliability. κ_w values of 0.40 or less suggest poor agreement, between 0.40 and 0.60 moderate, between 0.60 and 0.80 substantial and higher than 0.80 almost perfect agreement [13, 14, 29]. As per the ICC, values lower than 0.50 are considered poor, between 0.50 and 0.75 moderate, between 0.75 and 0.90 good, and higher than 0.90 excellent [30].

The aforementioned parameters were calculated for the total of assessments and then in subgroups, according to age, severity and diagnosis. To estimate the level of agreement in those cases where the verbal component of the scale could be reliably assessed, additional calculations

for the GCS verbal and total score were done separately, after the pseudoscored patients were excluded. Specifically, intubated patients or those with a tracheostomy, where the GCS verbal component could not be assessed, were pseudoscored with the lowest possible verbal score of 1 and, thus, were not included in this subgroup analysis.

The intrarater reliability was not tested in our study. This was decided because, as already noted in the literature, it cannot be reliably evaluated. Specifically, there are doubts whether a rater can assess free of bias a patient after a short time period without being influenced by previous scoring [14, 31–33].

Cronbach's α values were calculated for internal consistency assessment. Values higher than 0.70 are generally accepted to suggest good internal consistency, since if α is too high it may indicate item redundancy [34, 35].

For the evaluation of construct validity, we stated a priori hypotheses regarding the relationship between the FOUR score and other instruments measuring similar constructs. The GCS [2, 3] is the widely used performance-based observational clinical tool for the assessment of coma [4–6], with two of its three components (namely eye opening and motor response) related to those of the FOUR score [14]. The correlation between the FOUR score and the GCS (in terms of Spearman's ρ) has been repeatedly used as a measure of the former's construct validity [13]. In the literature, the translation of the FOUR score into other languages has resulted in Spearman's ρ values higher than 0.80 [36–38]. Thus, Spearman's ρ values indicating a high positive correlation between the two scales were expected for all four raters.

A power analysis was performed to select the minimum number of participants in line with the design of our study that would reach an adequate statistical strength. It was found that 84 subjects would reach a power of 80%, which was considered appropriate for the purposes of the current study. Previous similar single-center studies presented with a median sample of 87 patients and a median number of performed assessments of 168 [13]. P values below 0.05 were considered statistically significant.

The software packages SPSS version 25 (IBM Corporation, Armonk, NY), G*Power 3.1 (Heinrich-Heine-Universität Düsseldorf) and GraphPad Prism® 7 (GraphPad Software Incorporation, San Diego, CA) were used for the analysis.

Results

Development of the Greek Version

As defined by the World Health Organization [22], emphasis was given to conceptual equivalence. Minimal linguistic issues during the process were resolved unanimously by the expert panel after discussion. The

third version of the translated scale was unanimously accepted by the members of the expert panel and its back-translation was approved (Supplementary Material 1). The “Instructions for the Assessment of the Individual Categories of the FOUR” [14] were also translated in Greek. Booklets containing the scales of

interest with detailed application instructions were designed.

Validation in Neurosurgical Patients

Participants’ Characteristics and Assessments’ Data

A total of 102 rating sessions were performed on 99 patients, resulting in 408 assessments (Fig. 1). Sixty-four

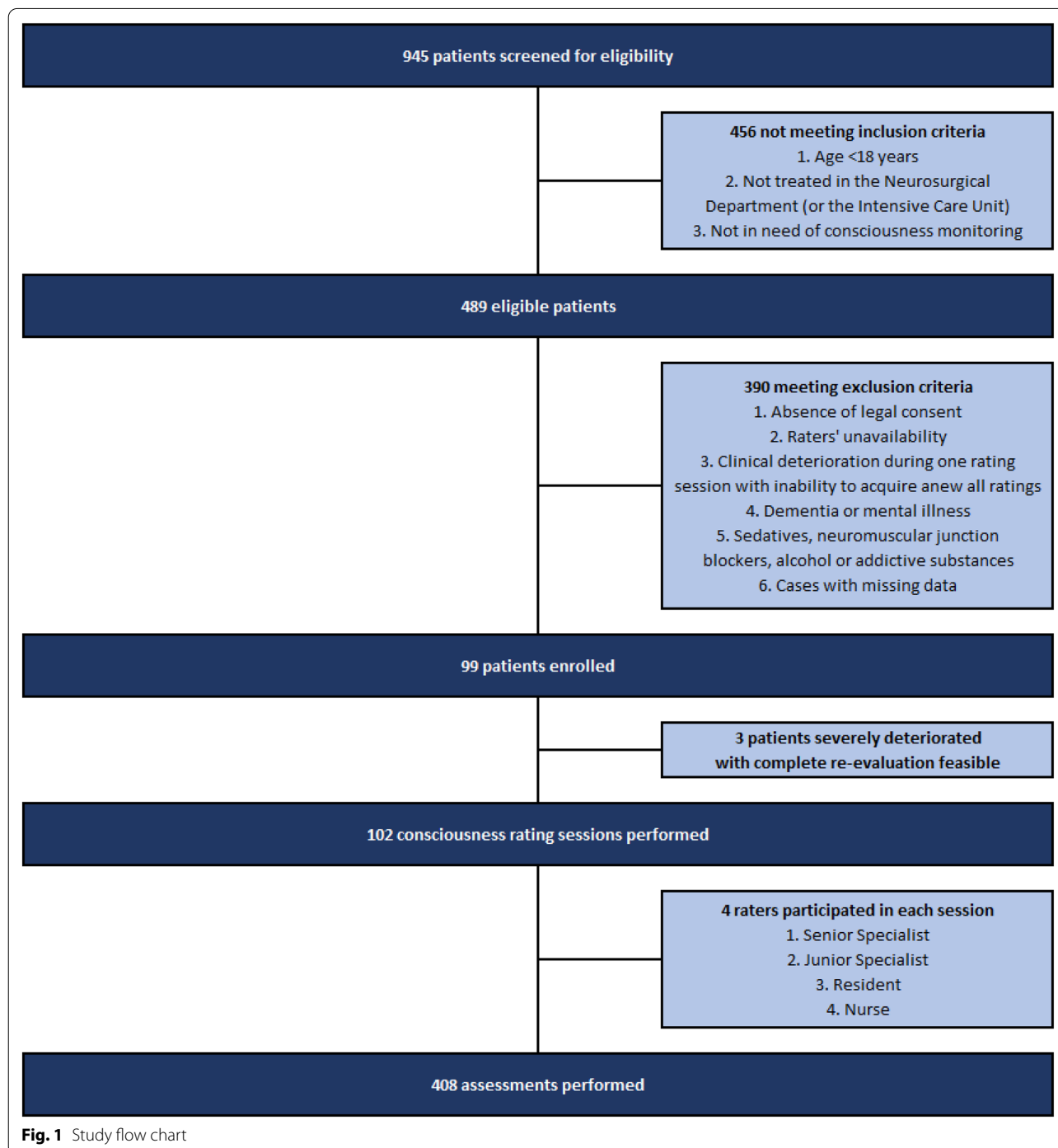
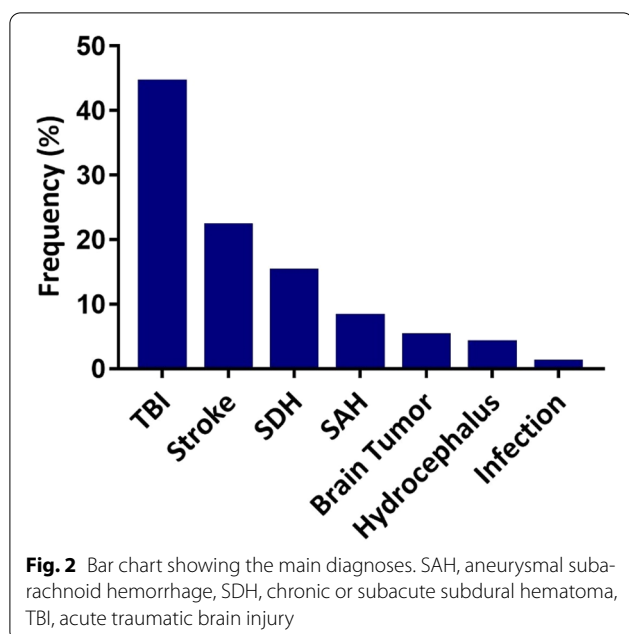


Fig. 1 Study flow chart



men and 35 women with a median age of 70 years (range 18–97) participated in the study. Their main diagnoses (Fig. 2) were acute traumatic brain injury (44 cases, 44.4%), ischemic or hemorrhagic stroke (22 cases, 22.2%), chronic or subacute subdural hematoma (15 cases, 15.2%), aneurysmal subarachnoid hemorrhage (8 cases, 8.1%), brain tumor (5 cases, 5.1%), hydrocephalus (4 cases, 4%), and central nervous system infection (1 case, 1%). Twenty-nine sessions (116 assessments, 28.4%) were performed on intubated patients or on patients with a tracheostomy tube. Other factors that would make the assessment of any of the scales' parameters impossible,

such as language and communication problems, tongue, and ocular trauma [11], were not recorded. No patients were excluded due to missing data.

According to the evaluations, the median GCS score was 10 for the two specialists and the nurses and 9.5 for the residents, whereas it was 13 for all raters' categories with the FOUR score. The minimum value of 3 for the GCS scored 28 times and the counterpart values of the FOUR score ranged from 3 to 8. FOUR score values between 0 and 2 were not recorded. This range of FOUR score values up to 8 for the most critical cases suggests its potential usefulness in such patients, the clinical condition of which cannot be described and monitored in detail by solely applying the GCS. On the other hand, the FOUR score's best possible value of 16 scored 105 times among patients who were not pseudoscoring. Interestingly, in 69.5% of those cases the counterpart values of the GCS were lower than 15 (range 12–14). The maximum value of the GCS was recorded 32 times. The counterpart values of the FOUR score were different than 16 in only 3 cases (15 in all of them). Those results suggest that the GCS probably has a similar advantage when assessing patients with mild consciousness disturbances.

Interrater Reliability

The κ_w values for the total scores were higher than 0.90 for all pairs of raters, suggesting almost perfect agreement for both scales. The lowest level of agreement was seen in the less experienced raters (resident and nurse), yet remained remarkably high (0.920 for the GCS, 0.918 for the FOUR score). The ICC values also indicated excellent overall agreement for the total scores (Tables 2 and 3).

Table 2 Quadratic weighted kappa (for each pair of raters) and intraclass correlation coefficient (for all raters) values for the GCS

Raters' pair	Total (95% CI)	E	V	M	T ^e	V ^e
Weighted kappa						
Sr specialist and Jr specialist	0.931 ^a (0.901–0.961)	0.794 ^b	0.920 ^a	0.828 ^a	0.937 ^a	0.878 ^a
Sr specialist and resident	0.940 ^a (0.913–0.966)	0.803 ^a	0.901 ^a	0.848 ^a	0.939 ^a	0.856 ^a
Sr specialist and nurse	0.934 ^a (0.903–0.965)	0.852 ^a	0.914 ^a	0.842 ^a	0.919 ^a	0.874 ^a
Jr specialist and resident	0.936 ^a (0.908–0.965)	0.842 ^a	0.898 ^a	0.873 ^a	0.926 ^a	0.849 ^a
Jr specialist and nurse	0.952 ^a (0.929–0.975)	0.884 ^a	0.915 ^a	0.869 ^a	0.951 ^a	0.871 ^a
Resident and nurse	0.920 ^a (0.879–0.961)	0.851 ^a	0.883 ^a	0.753 ^b	0.915 ^a	0.830 ^a
Overall ICC	0.936 ^a (0.914–0.953)	0.843 ^b	0.907 ^a	0.842 ^b	0.932 ^a	0.863 ^b

The GCS showed excellent interrater agreement for the total score and its three components in almost all cases. Values did not change significantly even after the exclusion of pseudoscoring patients

CI, Confidence Interval, E, Eye component, GCS, Glasgow Coma Scale, ICC, Intraclass Correlation Coefficient, M, Motor component, T^e, Total score with pseudoscoring patients excluded, V, Verbal component, V^e, Verbal component with pseudoscoring patients excluded

^a Denotes excellent agreement

^b Denotes good agreement

Table 3 Quadratic weighted kappa (for each pair of raters) and intraclass correlation coefficient (for all raters) values for the FOUR score

Raters' pair	Total (95% CI)	E	M	B	R
Weighted kappa					
Sr specialist and Jr specialist	0.950 ^a (0.930–0.971)	0.885 ^a	0.877 ^a	0.747 ^b	0.727 ^b
Sr specialist and resident	0.923 ^a (0.894–0.952)	0.847 ^a	0.870 ^a	0.647 ^b	0.618 ^b
Sr specialist and nurse	0.955 ^a (0.938–0.973)	0.893 ^a	0.859 ^a	0.823 ^a	0.812 ^a
Jr specialist and resident	0.929 ^a (0.901–0.958)	0.859 ^a	0.873 ^a	0.837 ^a	0.663 ^b
Jr specialist and nurse	0.964 ^a (0.945–0.982)	0.917 ^a	0.922 ^a	0.860 ^a	0.858 ^a
Resident and nurse	0.918 ^a (0.886–0.949)	0.852 ^a	0.820 ^a	0.776 ^b	0.689 ^b
Overall ICC	0.941 ^a (0.921–0.957)	0.879 ^b	0.875 ^b	0.784 ^b	0.733 ^c

Excellent interrater agreement was noted in most cases for the FOUR score and its components. The brainstem and respiration elements scored the lowest values; however, agreement remained at least moderate for all occasions

B, Brainstem component, CI, Confidence Interval, E, Eye component, FOUR, Full Outline of Unresponsiveness, ICC, Intraclass Correlation Coefficient, M, Motor component, R, Respiration component

^a Denotes excellent agreement

^b Denotes good agreement

^c Denotes moderate agreement

The values of κ_w for each one of the scales' components were higher than 0.60, indicating at least substantial agreement in all cases and for all pairs of raters; in most cases they were higher than 0.80, indicating excellent agreement. The lowest level of agreement was observed in the assessment of brainstem and respiratory components of the FOUR score, but remained high (0.647–0.860 and 0.618–0.858, respectively). The ICC assessing the overall agreement among the four raters for each component also suggested good or excellent agreement (ICC > 0.75), for almost all pairs of raters. As with the κ_w , the brainstem and respiratory components of the FOUR score showed the lowest level of agreement (Tables 2 and 3).

The interrater agreement was found to be comparable for the two scales by diagnosis group (acute trauma vs. other diagnoses) and age group (younger vs. older than the median age of 70 years), and excellent for all pairs of raters (>0.90). When grouping the patients to comatose and non-comatose (GCS score ≤ 8 vs. >8), there was a trend toward higher FOUR score values in cases with more severe disturbance of consciousness, remaining yet >0.60 on all occasions (Table 4). After excluding 29 cases that were pseudoscored for the GCS verbal component, the values remained higher than 0.80 in all cases, suggesting good to excellent agreement (Table 2).

Internal Consistency

The values of Cronbach's α for the GCS were 0.815, 0.828, 0.807, and 0.852 for the four raters (senior and junior specialist, resident, and nurse, respectively). The corresponding values for the FOUR score were 0.782, 0.724,

0.730, and 0.735, respectively, indicating good internal consistency for both scales.

Construct Validity

The values of Spearman's ρ between the FOUR score and the GCS were 0.936, 0.948, 0.906, and 0.930 for the four raters (senior and junior specialist, resident, and nurse, respectively). All correlations were found to be significant ($p < 0.001$; Fig. 3). Thus, the hypotheses of a high correlation between the two scales for all raters were confirmed, suggesting high construct validity for the Greek version of the FOUR score.

Discussion

In the present study, the Greek version of the FOUR score was presented and validated. Our results showed high interrater reliability, internal consistency, and construct validity for the FOUR score among raters. To our knowledge, this is the first study that provides such an in-depth assessment of interrater agreement when the scale is applied by raters with different levels of training and experience.

The possible impact of the raters' training and experience on interrater reliability has emerged since the initial validation of the FOUR score, where a lower agreement between nurses for some categories, in particular eye and brainstem components, was seen [14]. However, interrater reliability among nurses significantly improved in later studies [17]. Proper training and familiarity with the FOUR score were considered important factors in improving the agreement between raters, while the

Table 4 Quadratic weighted kappa (for each pair of raters) and intraclass correlation coefficient (for all raters) values for the two scales under assessment in patients with head trauma or not, in comatose and non-comatose patients, and in patients younger and older than 70 years

Categories	Diagnosis				Severity				Age			
	Acute trauma		Other		GCS ≤ 8		GCS > 8		≤ 70 years		> 70 years	
N	47		55		39		63		51		51	
Weighted kappa												
Raters' pair	GCS	FOUR	GCS	FOUR	GCS	FOUR	GCS	FOUR	GCS	FOUR	GCS	FOUR
Sr specialist and Jr specialist	0.948 ^a	0.945 ^a	0.917 ^a	0.954 ^a	0.704 ^b	0.865 ^a	0.771 ^b	0.802 ^a	0.921 ^a	0.958 ^a	0.940 ^a	0.943 ^a
Sr specialist and resident	0.939 ^a	0.915 ^a	0.939 ^a	0.930 ^a	0.799 ^b	0.745 ^b	0.792 ^b	0.736 ^b	0.925 ^a	0.923 ^a	0.953 ^a	0.924 ^a
Sr specialist and nurse	0.923 ^a	0.950 ^a	0.943 ^a	0.960 ^a	0.678 ^b	0.880 ^a	0.786 ^b	0.798 ^b	0.938 ^a	0.952 ^a	0.931 ^a	0.958 ^a
Jr specialist and resident	0.944 ^a	0.928 ^a	0.930 ^a	0.931 ^a	0.730 ^b	0.758 ^b	0.833 ^a	0.800 ^b	0.927 ^a	0.925 ^a	0.944 ^a	0.934 ^a
Jr specialist and nurse	0.962 ^a	0.955 ^a	0.944 ^a	0.970 ^a	0.786 ^b	0.883 ^a	0.864 ^a	0.880 ^a	0.941 ^a	0.952 ^a	0.962 ^a	0.975 ^a
Resident and nurse	0.919 ^a	0.905 ^a	0.920 ^a	0.927 ^a	0.618 ^b	0.729 ^b	0.796 ^b	0.717 ^b	0.898 ^a	0.915 ^a	0.941 ^a	0.920 ^a
Overall ICC	0.940 ^a	0.936 ^a	0.933 ^a	0.946 ^a	0.723 ^c	0.813 ^b	0.813 ^b	0.795 ^b	0.926 ^a	0.938 ^a	0.946 ^a	0.947 ^a

The interrater agreement remained at least moderate with no exceptions, ranging from good to excellent on most occasions. There was almost perfect agreement regardless of the diagnosis and age group. Note a trend toward a higher agreement in favor of the FOUR score when assessing comatose patients

FOUR, Full Outline of Unresponsiveness, GCS, Glasgow Coma Scale, ICC, Intraclass Correlation Coefficient, N, Number of cases

- ^a Denotes excellent agreement
- ^b Denotes good agreement
- ^c Denotes moderate agreement

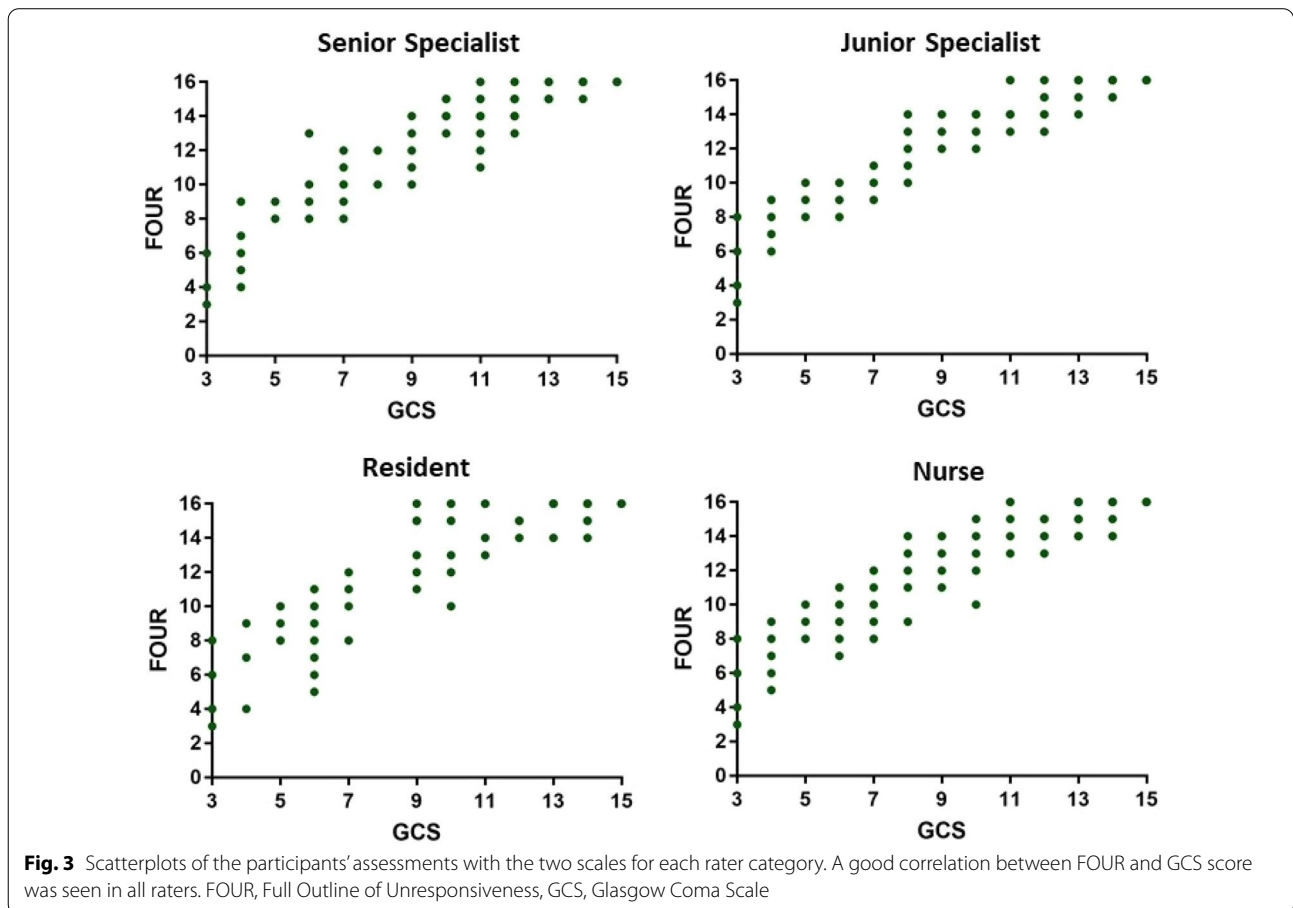


Fig. 3 Scatterplots of the participants' assessments with the two scales for each rater category. A good correlation between FOUR and GCS score was seen in all raters. FOUR, Full Outline of Unresponsiveness, GCS, Glasgow Coma Scale

different level of experience between nurses did not seem to play any important role [17].

According to our results, on most occasions, κ_w and ICC values showed good to excellent agreement between different raters, for both GCS and FOUR score, in all pairs of raters and all subgroups of patients, regardless of age, diagnosis or severity of consciousness disturbance. Therefore, it can be stated that the FOUR score can be reliably applied after proper training, despite the rater's level of education and experience. These findings agree with other reports from translated versions of the scale into various languages [36–39].

Even though medical practitioners in our hospital are widely experienced in the use of the GCS, the current results suggest comparable reliability between the two scales. The level of agreement remained good even among the less experienced raters (nurses and residents) for the brainstem and respiratory components of the FOUR score, the most demanding aspects of the scale with varying reliability results in the literature [19, 32, 33, 36, 40–43]. These findings indicate that, although the application of the FOUR score requires a more detailed assessment of the patient, a brief but vigorous training is sufficient to achieve an effective and reliable use in clinical practice. This is in line with the existing literature [31, 32]. Although the translation of a clinical tool in Greek was demanding due to the peculiarities of the language (for instance loanwords or repatriated words derived from the Greek language, especially in medical terminology), its feasibility and applicability in clinical practice were not influenced.

The absence of a verbal component has been presented as a potential advantage of the FOUR score over GCS, as it cannot be reliably assessed in many neurological patients [14]. Furthermore, the GCS has reportedly shown varying levels of interrater agreement, with some studies reporting κ_w values even less than 0.4, suggesting poor reliability [1]. Our results do not support this claim, since it was found to present excellent reliability among raters even after the exclusion of pseudoscored patients (untestable verbal component misleadingly increases interrater agreement, since all raters pseudoscore it with 1), for both the total score and verbal component.

Another reported advantage of the FOUR score is that, in the most severe cases, the components evaluating brainstem and respiration functions add useful neurological information, which allow for further subcategorization of individuals presenting with the lowest possible GCS score [13, 14]. This observation was clearly confirmed by the current results. It should be also mentioned that the FOUR score's interrater reliability when assessing comatose patients was found to

be higher compared with GCS for most pairs of raters (except for the senior specialist vs. resident), suggesting a potential superiority of the FOUR score in this subgroup of patients.

On the other hand, based on the present findings, it can be supported that cases with the less disturbed level of consciousness which present with the best possible FOUR score can be further subcategorized using the GCS. This is an expected finding, as the verbal component in the FOUR score is absent and patients with mild disturbance of consciousness often show confused communication. Further, it can be suggested that omitting the assessment of verbal communication deprives crucial information for cases that present with mild, or even moderate, disturbance of consciousness. This constitutes the majority of patients treated in neurosurgical and neurological wards, who need close monitoring for early detection of clinical deterioration [15, 16]. The verbal assessment retains clinical importance in mild cases usually hospitalized in neurosurgical and neurological departments, as these patients rarely present with disturbed brainstem function and respiratory pattern, therefore a possible replacement of the GCS cannot be supported. This is not the case, however, for patients with severe unresponsiveness resulting in inability to interact verbally, where the FOUR score shows undeniable advantages.

According to the current results, both GCS and FOUR score present with high interrater reliability and internal consistency, but also correlate significantly with each other. However, it cannot be suggested that differences between them are insignificant and the selection of one of them in clinical practice is usually a matter of personal or institutional preference. Nevertheless, each scale presents with potential advantages when applied in patients with specific characteristics.

This study has some limitations. It is a single-center study, limited to patients of neurosurgical concern. It has to be noted, however, that all were highly in need of accurate level of consciousness assessment and monitoring, with a notable number of them presenting with severely disturbed level of consciousness. Furthermore, the number of patients under mechanical ventilation that were included was limited. Patients under sedation were also excluded. Thus, our results need further verification for those and, since the FOUR score's advantages are likely associated with patients in a more severe condition, future studies should probably focus on this category. Except for the two specialists, it was not feasible to obtain four constant raters. Although this was clearly the case in all previous studies, it might have affected the ratings' consistency. However, all raters underwent similar intensive training for the application of the scales,

therefore minimizing the risk of bias. To avoid unnecessary disruptions and inconveniences during the study, a translation of the GCS through a similar vigorous procedure was not performed, since the existing Greek version has been widely and successfully used for many years. Finally, despite the meticulous prospective design, slight clinical deterioration that remained unnoticed, influencing thus the level of consciousness between ratings, cannot be ruled out.

Nevertheless, this study presents one of the largest numbers of total ratings for the assessment of interrater reliability. It is the first in the Greek population and it was designed to include a large number of raters, both experienced and less experienced, also nurses who were not well familiar with the neurological examination. Furthermore, contrary to previously published reports in which patients were divided into groups each examined by a different pair of raters, all participants were evaluated by all categories of examiners. We were able to successfully evaluate the agreement among health care professionals with different levels of experience and training, highlighting, thus, the effectiveness and simplicity of training necessary for the reliable application of the FOUR score.

Conclusions

The Greek version of the FOUR score is a valid and reliable tool for assessing the level of consciousness. It has high interrater reliability, internal consistency and construct validity, which is comparable with GCS, with similar observations among raters with different training and experience. Its clinical importance is probably greater for patients with severely disturbed level of consciousness. Studies including larger number of patients under mechanical ventilation are needed to further validate the present findings.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1007/s12028-021-01342-w>.

Author details

¹ Department of Neurosurgery, Hippokraton General Hospital, Aristotle University School of Medicine, 49 Konstantinoupoleos Str., 54642 Thessaloniki, Greece. ² Department of Neurosurgery, AHEPA University Hospital, Aristotle University School of Medicine, Thessaloniki, Greece. ³ Greek Association of Alzheimer's Disease and Related Disorders, Thessaloniki, Greece.

Acknowledgements

The authors need to express their gratitude to Professor Eelco FM Wijdicks, for the permission to translate the Full Outline of Unresponsiveness Score in Greek and for his approval of the final version, to Ioannis Gounaris, for his contribution in the back-translation, and to Annika Hickisch, registered nurse at Örebro University Hospital, Sweden, for providing us with valuable information and useful discussions. Furthermore, to neurosurgery residents Dimitrios Tsirikis, Nikolaos Karanikolas, Panagiotis Varoutis, and Panagiotis Monioudis and nurses Eleni Pappa, Anna Mourouglaki, Evangelos Tsangarakis, Panagiotis

Papamanolis, Maria Vordou, Periklis Giaglis, Eferpi Kosmidou, and Anna Vradeli, for their invaluable contribution in patients' assessments.

Authors' contributions

DMA contributed to the conception and design of the work, the acquisition and evaluation of data and the drafting of the article. PPT contributed to the conception and design of the work, the evaluation of data, and the critical revision of the content. NGF contributed to the critical revision of the content. MST contributed to the design of the work and the critical revision of the content. KM and MT contributed to the acquisition of the data. CAT contributed to the conception of the work and the critical revision of the content. All authors were involved in the final approval of the manuscript.

Source of support

All authors confirm that this study did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

All authors declare that they have no conflicts of interest.

Ethical approval/informed consent

The authors confirm adherence to ethical guidelines. The study protocol was approved by the ethics committee of the hospital (Ref. Nr. 985-2017). In compliance with the current legislation, the National Data Protection Authority was notified on its conduction (Ref. Nr. 850-2018). The study was conducted in accordance with the ethical standards of the 1964 Helsinki Declaration and its later amendments. Informed consent was obtained from all individual participants included in the study.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 25 May 2021 Accepted: 26 August 2021

Published: 23 September 2021

References

- Kornbluth J, Bhardwaj A. Evaluation of coma: a critical appraisal of popular scoring systems. *Neurocrit Care*. 2011;14(1):134–43.
- Teasdale G, Jennett B. Assessment of coma and impaired consciousness. A practical scale. *Lancet*. 1974;2(7872):81–4.
- Teasdale G, Jennett B. Assessment and prognosis of coma after head injury. *Acta Neurochir (Wien)*. 1976;34(1–4):45–55.
- Sozzi M, Inzaghi MG. Instruments for evaluation of altered states of consciousness. *Neuropsychol Trends*. 2011;10:25–42.
- Moore SA, Wijdicks EF. The acutely comatose patient: clinical approach and diagnosis. *Semin Neurol*. 2013;33(2):110–20.
- Idrovo Freire LA. The FOUR score: is it just another new coma scale? *Intern Emerg Med*. 2012;7(3):203–4.
- Laureys S, Bodart O, Gosseries O. The Glasgow Coma Scale: time for critical reappraisal? *Lancet Neurol*. 2014;13(8):755–7.
- Matis G, Birbilis T. The Glasgow coma scale—a brief review. Past, present, future. *Acta Neurol Belg*. 2008;108(3):75–89.
- Prasad K. The Glasgow Coma Scale: a critical appraisal of its clinimetric properties. *J Clin Epidemiol*. 1996;49(7):755–63.
- Sternbach GL. The Glasgow coma scale. *J Emerg Med*. 2000;19(1):67–71.
- Middleton PM. Practical use of the Glasgow Coma Scale; a comprehensive narrative review of GCS methodology. *Australas Emerg Nurs J*. 2012;15(3):170–83.
- Healey C, Osler TM, Rogers FB, et al. Improving the Glasgow Coma Scale score: motor score alone is a better predictor. *J Trauma*. 2003;54(4):671–8.
- Anestis DM, Tsitsopoulos PP, Tsonidis CA, Foroglou N. The current significance of the FOUR score: a systematic review and critical analysis of the literature. *J Neurol Sci*. 2020;409:116600.
- Wijdicks EF, Bamlet WR, Maramattom BV, Manno EM, McClelland RL. Validation of a new coma scale: the FOUR score. *Ann Neurol*. 2005;58(4):585–93.
- Servadei F. Coma scales. *Lancet*. 2006;367(9510):548–9.

16. Stevens RD, Sutter R. Prognosis in severe brain injury. *Crit Care Med*. 2013;41(4):1104–23.
17. Wolf CA, Wijdicks EF, Bamlet WR, McClelland RL. Further validation of the FOUR score coma scale by intensive care nurses. *Mayo Clin Proc*. 2007;82(4):435–8.
18. Almojuela A, Hasen M, Zeiler FA. The Full Outline of UnResponsiveness (FOUR) Score and its use in outcome prediction: a scoping systematic review of the adult literature. *Neurocrit Care*. 2018;31(1):162–75.
19. Fischer M, Ruegg S, Czaplinski A, et al. Inter-rater reliability of the Full Outline of UnResponsiveness score and the Glasgow Coma Scale in critically ill patients: a prospective observational study. *Crit Care*. 2010;14(2):R64.
20. Kevric J, Jelinek GA, Knott J, Weiland TJ. Validation of the Full Outline of Unresponsiveness (FOUR) Scale for conscious state in the emergency department: comparison against the Glasgow Coma Scale. *Emerg Med J*. 2011;28(6):486–90.
21. Kramer AA, Wijdicks EF, Snively VL, et al. A multicenter prospective study of interobserver agreement using the full outline of unresponsiveness score coma scale in the intensive care unit. *Crit Care Med*. 2012;40(9):2671–6.
22. World Health Organization. Process of translation and adaptation of instruments. https://www.who.int/substance_abuse/research_tools/translation/en/.
23. Institute of Neurological Sciences NHS Greater Glasgow and Clyde. The Glasgow coma scale and application guidelines (Greek). <https://www.glasgowcomascale.org/downloads/GCS-Assessment-Aid-Greek.pdf>.
24. Von Elm E, Altman DG, Egger M, et al. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *Ann Intern Med*. 2007;147(8):573–7.
25. Gagnier JJ, Lai J, Mokkink LB, Terwee CB. COSMIN reporting guideline for studies on measurement properties of patient-reported outcome measures. *Qual Life Res*. 2021;30(8):2197–218.
26. Mokkink LB, Terwee CB, Knol DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. *BMC Med Res Methodol*. 2010;10(1):1–8.
27. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol*. 2010;63(7):737–45.
28. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Qual Life Res*. 2010;19(4):539–49.
29. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977:159–74.
30. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15(2):155–63.
31. Akavipat P. Endorsement of the FOUR score for consciousness assessment in neurosurgical patients. *Neurol Med Chir (Tokyo)*. 2009;49(12):565–71.
32. Stead LG, Wijdicks EF, Bhagra A, et al. Validation of a new coma scale, the FOUR score, in the emergency department. *Neurocrit Care*. 2009;10(1):50–4.
33. Matheesiriwat N, Kuptniratsaikul S. The FOUR score and Glasgow Coma Scale to evaluate the patients with intubation at emergency room. *Royal Thai Army Med J*. 2012;65(3):145–52.
34. Bland JM, Altman DG. Statistics notes: Cronbach's alpha. *BMJ*. 1997;314(7080):572.
35. Tavakol M, Dennick R. Making sense of Cronbach's alpha. *Int J Med Educ*. 2011;2:53.
36. Idrovo L, Fuentes B, Medina J, et al. Validation of the FOUR Score (Spanish Version) in acute stroke: an interobserver variability study. *Eur Neurol*. 2010;63(6):364–9.
37. Marcati E, Ricci S, Casalena A, et al. Validation of the Italian version of a new coma scale: the FOUR score. *Intern Emerg Med*. 2012;7(2):145–52.
38. Peng J, Deng Y, Chen F, et al. Validation of the Chinese version of the FOUR score in the assessment of neurosurgical patients with different level of consciousness. *BMC Neurol*. 2015;15:254.
39. Hickisch A, Holmefur M. Swedish translation and reliability of the full outline of unresponsiveness score. *J Neurosci Nurs*. 2016;48(4):195–205.
40. Bruno MA, Ledoux D, Lambermont B, et al. Comparison of the full outline of unresponsiveness and Glasgow liege Scale/Glasgow Coma Scale in an intensive care unit population. *Neurocrit Care*. 2011;15(3):447–53.
41. Gujjar AR, Jacob PC, Nandhagopal R, et al. Full outline of un responsiveness score and Glasgow Coma Scale in medical patients with altered sensorium: interrater reliability and relation to outcome. *J Crit Care*. 2013;28(3):316.E1–8.
42. Mercy A, Thakur SR, Yaddanapudi S, Bhagat H. Can FOUR Score replace GCS for assessing neurological status of critically ill patients-An Indian Study. *Nurs Midwifery Res J*. 2013;9(2):63–72.
43. Lee TKP, Kitchell AKB, Siu AYC, Chen NK. Validation of the Full Outline of Unresponsiveness score coma scale in patients clinically suspected to have acute stroke in the emergency department. *Hong Kong Journal of Emergency Medicine*. 2017;24(5):230–6.