Witness Statement Ref. No. 178/3

NAME OF CHILD: Claire Roberts

Name: Ian Young

Title: Professor

Present position and institution:

Professor of Medicine and Director of The Centre for Public Health, Queen's University Belfast

Previous position and institution:
[As at the time of the child's death]

Membership of Advisory Panels and Committees:
[Identify by date and title all of those between January 1995-August 2012]

Previous Statements, Depositions and Reports:
[Identify by date and title all those made in relation to the child's death]

OFFICIAL USE:
List of previous statements, depositions and reports attached:

Ref:	**Date:**	
096-007-039		Statement to the PSNI
091-010-060	4th May 2006	Deposition to the Coroner
WS-178/1	14th Sep 2012	Inquiry Witness Statement
WS-178/2	30th Sep 2012	Inquiry Witness Statement

Comments on the interpretation of changes in the Glasgow Coma Scale of Claire Roberts during 22nd/23rd October.

Professor Ian S.Young

The purpose of this report is to address the fluctuations in Claire's GCS during her admission to Allen Ward, and to comment on the interpretation of these. I wish to highlight significant issues around the interpretation of GCS scores which have not, to my knowledge, been identified by any of the witnesses to date. The inquiry has in particular focussed on the fall in GCS from 8 at 8pm to 6 at 9pm, and a number of witnesses have been asked to agree that this represents deterioration in Claire's condition, which they have generally accepted. These witnesses are clearly not aware of the significant literature about measurement variability in GCS assessment, and it is this and its consequences which I wish to draw to the attention of the Inquiry.

In addition, a number of expert witnesses appear unaware of this literature and its relevance. For example, Dr. MacFaul says at **238-002-075**: *"It was stated that Claire's CNS observations had remained stable over a period of time and no clinical signs of further deterioration were noted. This is not correct, the GCS reduced over the evening and had done so by the time the blood sodium level was available."* In view of the evidence which I summarise below, this statement is unreliable. The GCS values during the day are entirely compatible with Claire's neurological condition remaining stable, though she was clearly seriously ill, and should not be interpreted as indicating a decline in her condition over this period.

Clinical history and relevant evidence:

Dr.MacFaul provides a useful table documenting GCS scores during the day at **238-002-219.** There was initial reading of 9 at 1pm and subsequent values fluctuated between 8 and 6. The inquiry has in particular focussed on the fall in GCS from 8 at

8pm to 6 at 9pm.
During the course of the day, CNS observations were made by several different nurses, who then proceeded to calculate and record the GCS. Of note, the CNS observations at 8pm and 9pm and calculation of the GCS appear to have been done by different nurses, due to the change over in shifts at around that time.

A significant body of scientific evidence shows that there is considerable variability in assessment of GCS and related scales by different observers. If two different assessors examine a patient and assess the GCS they may come up with significantly different values. This is referred to as inter-observer variability. In the case of the GCS, it is perfectly possible for two observers independently assessing a single patient to come up with a GCS value differing by 2 or even more points due to inter-observer variability. This has led to attempts to develop alternative clinical scoring systems with improved reproducibility in recent years, although the GCS remains widely used.

A number of scientific studies have investigated and documented inter-observer variability in GCS assessment. Rowley and Fielding (Lancet 1991:227:535-8) described the “Reliability and accuracy of the Glasgow Coma Scale with experienced and inexperienced users”. They compared the performance of four different groups of nurses, classified according to experience, and assessed how accurately they calculated GCS value compared with an expert. They concluded that while the GCS was used accurately by experienced and highly trained users, inexperienced users made significant errors. The errors were substantial, averaging in some cases more than one point on the four-point and five point scales of the GCS.

Tatman et al. looked at a modification of the Glasgow Coma Scale in more detail in Archives of Disease in Childhood 1997;77:519–521. Two observers made observations within 15 minutes of each other. One observer was the patient’s nurse and the other a trained investigator. Inter-observer reliability was determined between the first and second observation for each component of the scale. Seventy three children had 104 sets of observations. The results are presented in a way which shows the magnitude of the differences in a way which is relatively easy to understand. Therefore, I reproduce Table 4 from the paper on the next page

.

The top row shows the modified GCS score calculated by the first observer. If you follow the column down then you see the scores recorded by the second observer. For example, taking 7 as the GCS score calculated by the first observer, the second observer examining the same stable patient within 15 minutes calculated the score as 6 (two occasions), 7 (two occasions), 8 (3 occasions) and 10 (once). Taking 8 as the score determined by the first observer, we see that the second score was 6 (once), 8 (twice), 9 (twice) or 10 (twice).

This paper clearly demonstrates that a two point change in GCS score may simply represent inter-observer variability, and is in keeping with the Lancet paper referred to above where the results are presented in a more technical fashion.

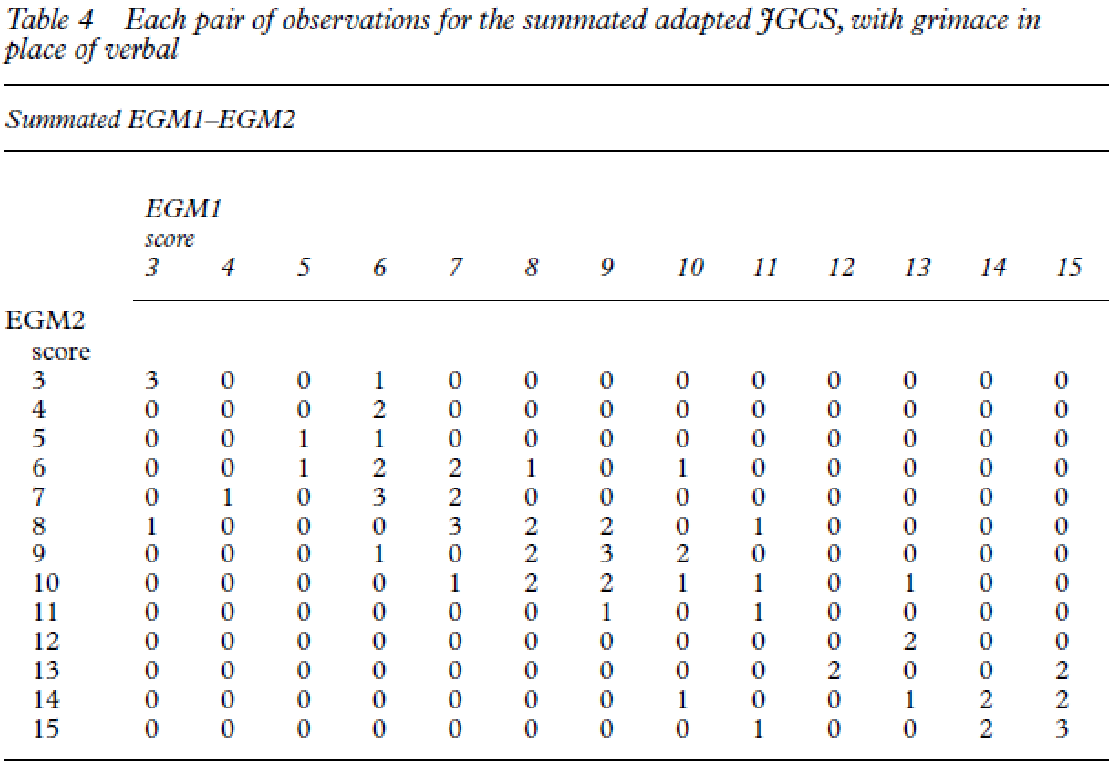

Table 4 Each pair of observations for the summated adapted JGCS, with grimace in place of verbal

Summated EGM1–EGM2

	EGM1 score												
	3	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
EGM2 score													
3	3	0	0	1	0	0	0	0	0	0	0	0	0
4	0	0	0	2	0	0	0	0	0	0	0	0	0
5	0	0	1	1	0	0	0	0	0	0	0	0	0
6	0	0	1	2	2	1	0	1	0	0	0	0	0
7	0	1	0	3	2	0	0	0	0	0	0	0	0
8	1	0	0	0	3	2	2	0	1	0	0	0	0
9	0	0	0	1	0	2	3	2	0	0	0	0	0
10	0	0	0	0	1	2	2	1	1	0	1	0	0
11	0	0	0	0	0	0	1	0	1	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	2	0	0
13	0	0	0	0	0	0	0	0	0	2	0	0	2
14	0	0	0	0	0	0	0	1	0	0	1	2	2
15	0	0	0	0	0	0	0	0	1	0	0	2	3

From Tatman et al., Archives of Disease in Childhood 1997;77:519–521.
Fischer et al. in *Critical Care 2010,**14**:R64* assessed the inter-rater reliability of the Glasgow Coma Scale in critically ill patients, looking at 267 consecutive adult patients

admitted to critical care. The observers were neurologists or ICU staff, and therefore likely to be relatively skilled. The two consecutive observers produced identical GCS scores on only 71% of occasions. Again, the results are presented in a format which allows the extent of discrepancies to be appreciated relatively easily.

Firstly, in the figure below we see in the left half paired results within one hour for patients assessed by two observers, neurologists in the top panel and ICU staff in the bottom panel. For neurologists, there were five patients assigned a GSC score of 7 by the first observer. The second neurologist scored them at 6,7,8,10 and 11. For ICU staff three patients were assigned a score of 7 by the first observer. The second scored them at 5, 7 and 9.

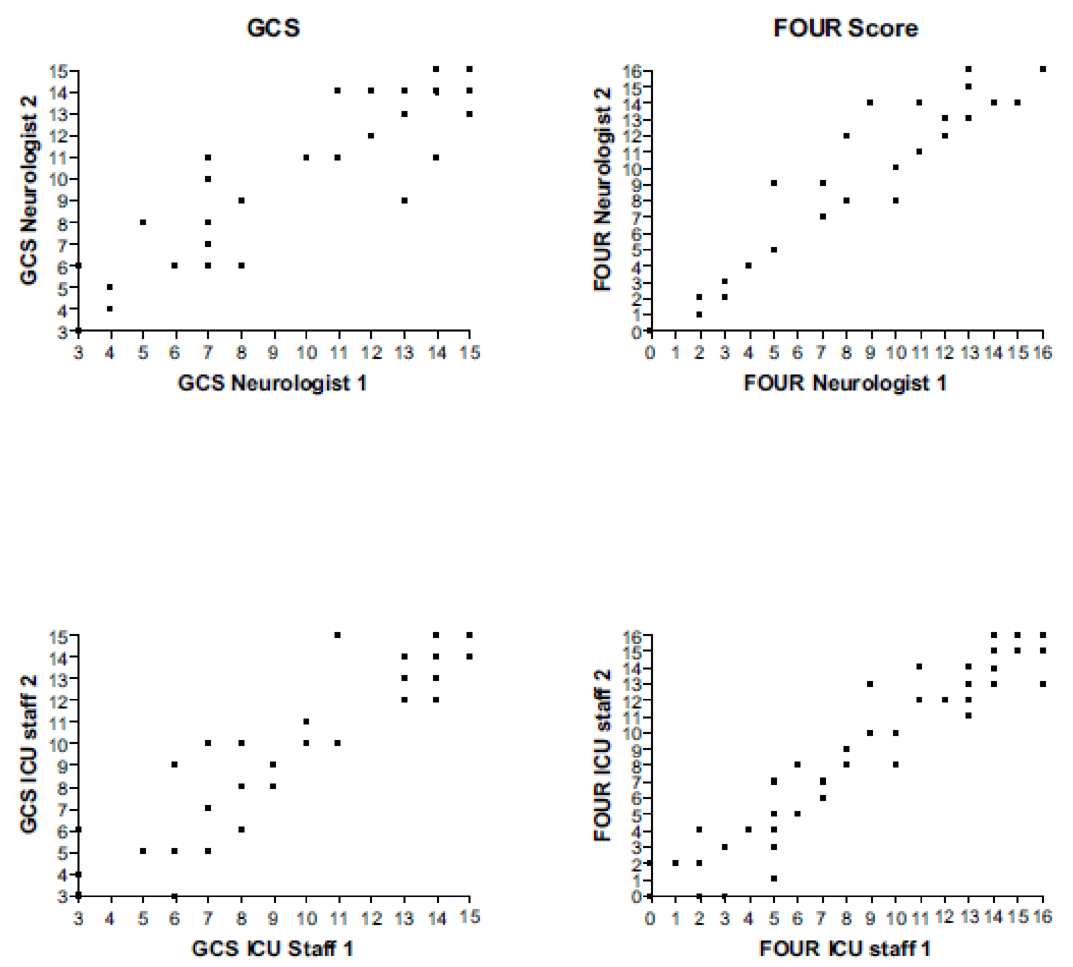

Figure 2 Inter-rater agreement of GCS and FOUR scores. Scatterplots of the pair-wise ratings of neurologists (top panels), ICU staff (middle panels), and neurologist-ICU staff (bottom panels) for the Glasgow Coma Scale (GCS; left side panels) and Full Outline of UnResponsiveness (FOUR) score (right side panels).

In the figure below (figure 3 from the paper), we see the extent of disagreement between different categories of staff (defined as a difference of greater than one point between them). Notably, the greatest level of disagreement was in the GCS 6-8 range, with 30 – 50% disagreement between different staff groups.

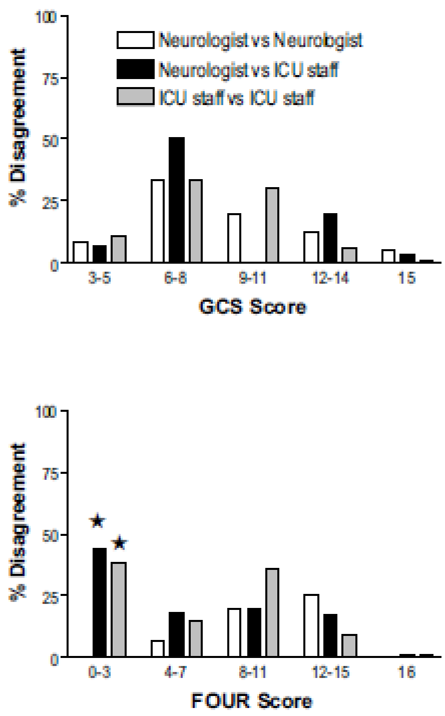

Figure 3 Disagreement rates for GCS and FOUR scores. Disagreements of more than one score point in pair-wise ratings of the Glasgow Coma Scale (GCS) score (top panel) and the Full Outline of UnResponsiveness (FOUR) score (bottom panel) respectively. Scores are divided into quartiles. As a substantial proportion of ratings were at the maximum of the each scale (i.e. GCS 15, FOUR 16), the maximum category is shown separately in addition to the quartiles. Disagreements are expressed as a percentage of the total number of ratings in a given quartile of the GCS score and FOUR score, respectively. White bars = disagreements between the neurologists; black bars = disagreements between the neurologists and ICU staff; grey bars = disagreements between ICU staff. For both scores, disagreements were significantly ($P <$ 0.001) less frequent in the maximum category (i.e. GCS 15, FOUR 16) than in all other categories.* For the lowest quartile of the FOUR score, the disagreement between neurologist and ICU staff ($P = 0.034$) and between ICU staff and ICU staff ($P = 0.045$) was significantly greater than that between the neurologists.

The results in the paper clearly show that a two point difference in GCS recorded by different individuals may simply be a result of inter-individual variability and cannot be viewed as a reliable indicator of clinical deterioration.

Conclusions:

This evidence is summarised above. In view of this evidence, it is simply not possible to say that a score of 6 assessed by one observer represents a change in condition when compared to a score of 8 assessed by a different observer. The variation in recorded

GCS values during the day is entirely compatible with Claire’s neurological condition remaining stable, though she was clearly seriously ill, and should not be interpreted as indicating a decline in her condition over this period.

At **238-002-075** Dr. MacFaul makes the following comment, criticizing information provided to Claire’s parents in 2004:

“362. It was stated that Claire‘s CNS observations had remained stable over a period of time and no clinical signs of further deterioration were noted. This is not correct, the GCS reduced over the evening and had done so by the time the blood sodium level was available.”

This statement takes no account of the impact of inter-observed variability in GCS assessment discussed above and in light of published scientific evidence is incorrect.

Professor Ian S.Young
2/11/12

Fischer *et al. Critical Care* 2010, **14**:R64
http://ccforum.com/content/14/2/R64

CRITICAL CARE

RESEARCH Open Access

Inter-rater reliability of the Full Outline of UnResponsiveness score and the Glasgow Coma Scale in critically ill patients: a prospective observational study

Michael Fischer[1], Stephan Rüegg[2], Adam Czaplinski[2], Monika Strohmeier[1], Angelika Lehmann[1], Franziska Tschan[3], Patrick R Hunziker[1] and Stephan C Marsch*[1]

Abstract

Introduction: The Glasgow Coma Scale (GCS) is the most widely used scoring system for comatose patients in intensive care. Limitations of the GCS include the impossibility to assess the verbal score in intubated or aphasic patients, and an inconsistent inter-rater reliability. The FOUR (Full Outline of UnResponsiveness) score, a new coma scale not reliant on verbal response, was recently proposed. The aim of the present study was to compare the inter-rater reliability of the GCS and the FOUR score among unselected patients in general critical care. A further aim was to compare the inter-rater reliability of neurologists with that of intensive care unit (ICU) staff.

Methods: In this prospective observational study, scoring of GCS and FOUR score was performed by neurologists and ICU staff on 267 consecutive patients admitted to intensive care.

Results: In a total of 437 pair wise ratings the exact inter-rater agreement for the GCS was 71%, and for the FOUR score 82% ($P = 0.0016$); the inter-rater agreement within a range of ± 1 score point for the GCS was 90%, and for the FOUR score 92% ($P =$ ns.). The exact inter-rater agreement among neurologists was superior to that among ICU staff for the FOUR score (87% vs. 79%, $P = 0.04$) but not for the GCS (73% vs. 73%). Neurologists and ICU staff did not significantly differ in the inter-rater agreement within a range of ± 1 score point for both GCS (88% vs. 93%) and the FOUR score (91% vs. 88%).

Conclusions: The FOUR score performed better than the GCS for exact inter-rater agreement, but not for the clinically more relevant agreement within the range of ± 1 score point. Though neurologists outperformed ICU staff with regard to exact inter-rater agreement, the inter-rater agreement of ICU staff within the clinically more relevant range of ± 1 score point equalled that of the neurologists. The small advantage in inter-rater reliability of the FOUR score is most likely insufficient to replace the GCS, a score with a long tradition in intensive care.

Introduction

The assessment of comatose patients is an important part of critical care. Unfortunately, there is no objective measure of coma like temperature or blood pressure. Thus, so far the assessment of the level of coma has to rely on clinical scores. The Glasgow Coma Scale (GCS), originally designed for patients with head trauma [1], has become the most widely used scoring system for patients with an altered level of consciousness in the ICU. Important limitations of the GCS include inconsistent inter-observer reliability [2], concerns over the predictive value in brain injury patients undergoing modern neuro-intensive care [3], the impossibility of assessing the verbal score in intubated patients, and the exclusion of brainstem reflexes. Over the past decades, a variety of alternative scoring systems have been developed [4-7], although none of them reached widespread acceptance.

* Correspondence: smarsch@uhbs.ch
[1] Department of Medical Intensive Care, University Hospital, Spitalstrasse, Basel, 4031, Switzerland
Full list of author information is available at the end of the article

The FOUR (Full Outline of UnResponsiveness) score, a coma scale consisting of four components (eye response, motor response, brainstem reflexes, and respiration pattern) was recently proposed by investigators from the Mayo Clinic [8]. Validation among patients receiving no sedative agents by dedicated staff in neuro-intensive care demonstrated good to excellent inter-rater reliability [8,9]. By contrast to the GCS, the FOUR score does not rely on a verbal response. In the ICU, a variety of conditions such as intubation, sedation, or delirium preclude a reliable assessment of a verbal response and, therefore, the FOUR score is an attractive tool. However, before this new score can be recommended for routine use in the ICU, the following limitations should be addressed: so far the FOUR score has not been validated in critically ill patients outside of the Mayo Clinic; so far the FOUR score has not been validated in sedated patients; and so far the FOUR score has only been validated by dedicated staff in neuro-ICUs. This may have resulted in a much higher inter-rater reliability than that achievable by ICU staff of general ICUs.

Accordingly, the aims of the present study were: to compare the inter-rater reliability of the GCS and the new FOUR score among unselected patients in general medical ICU; and to compare the inter-rater reliability of neurological scoring provided by staff members of general medical ICUs with that of neurologists.

Materials and methods

The study was performed on one of the two subunits of the medical ICU of the University Hospital of Basel, Switzerland. The study was approved by the regional ethical committee. As GCS scoring was already routinely performed on our unit prior to the study and no therapeutic decisions were based on the FOUR scoring, the ethical committee waived the need to obtain individual informed consent.

Ratings were performed by two board-certified staff neurologists (S.R. and A.C.) serving as gold standard, eight ICU nurses, and four ICU physicians. Prior to the study, all raters received an instruction by one of the neurologists including a supervised scoring of GCS and FOUR score in two patients.

We prospectively studied the FOUR score and the GCS in consecutive adult patients admitted to our ICU. Exclusion criteria were the unavailability of both neurologists and the patients' unwillingness to participate in the ratings. Scoring was performed between 9:00 am and 10:00 am on weekdays only. Scoring occurred at the first possible occasion after admittance and each patient was scored only once. Eligible patients were identified by the head-nurse and colour-coded on the main board showing all patients presently admitted. If available, raters performed their ratings on the coded patients in the time frame specified. Raters were not aware of other ratings or the results thereof. Patients were included if at least one of the neurologists and one member of the ICU staff were able to perform a rating within a time interval of one hour. In addition, 100 consecutive patients were rated by both of the two neurologists to assess their inter-rater agreement. Patients were included if both neurologists were able to perform their ratings within a time interval of one hour.

For GCS scoring, the raters used a one-sided A4-sized form containing written instructions. In intubated patients, the rating for the verbal domain of the GCS was defined to be 1. For the FOUR scoring, the raters used a one-sided A4-sized form containing both written and visual instruction: the written instruction was a German translation of the original instruction from the Mayo Clinic [8]; the visual instruction was a coloured copy of the version published in 2005 [8], adapted in size to fit the scoring form. The definition of the FOUR score and the GCS are displayed in Table 1.

Acute physiology and chronic health evaluation (APACHE) II scores were obtained for the first 24 hours after admittance to the ICU. For patients that stayed for 28 days or more in our hospital or died during their hospitalisation, 28-day mortality was assessed using the in-hospital electronic patient documentation system. Twenty eight-day mortality of discharged patients was assessed by contacting the physician treating the patient at home or in another institution.

Statistics

Data were analysed using SPSS (version 15.0), a commercially available statistical software. Three categories of pair-wise ratings were analysed: 1) neurologist - neurologist, 2) ICU staff ICU staff, and, 3) neurologist ICU staff. For each category no more than one pair-wise rating was analysed in every patient. In case of more than one pair-wise rating in a given category (e.g. patient was rated by two neurologists and two members of ICU staff resulting in four pair-wise ratings in the category neurologist ICU staff) the rating to be analysed was randomly chosen using computer-generated numbers. Pair-wise-weighted kappa values were calculated for the GCS and the FOUR score. A kappa value of 0.4 or less is considered poor, values between 0.4 and 0.6 are considered fair to moderate, values between 0.6 and 0.8 are considered good, and values above 0.8 are considered excellent agreement [10]. Although assessment of inter-rater reliability using kappa statistics is scientifically appropriate, this approach does not result in measures of obvious clinical usefulness. Rather than an exact agreement we determined that for the dynamic environment of the ICU a precision in scoring within the range of ± 1 score points for both GCS and FOUR score would be sufficient for the majority, if not all,

Table 1: Definition of the FOUR score and the Glascow Coma Score

FOUR score	Glascow Coma Scale
Eye response	Eye response
4 = eyelids open or opened, tracking, or blinking to command	4 = eyes open spontaneously
3 = eyelids open but not tracking	3 = eye opening to verbal command
2 = eyelids closed but open to loud voice	2 = eye opening to pain
1 = eyelids closed but open to pain	1 = no eye opening
0 = eyelids remain closed with pain	Motor response
Motor response	6 = obeys commands
4 = thumbs-up, fist, or peace sign	5 = localising pain
3 = localising to pain	4 = withdrawal from pain
2 = flexion response to pain	3 = flexion response to pain
1 = extension response to pain	2 = extension response to pain
0 = no response to pain or generalised myoclonus status	1 = no motor response
Brainstem reflexes	Verbal response
4 = pupil and corneal reflexes present	5 = oriented
3 = one pupil wide and fixed	4 = confused
2 = pupil or corneal reflexes absent	3 = inappropriate words
1 = pupil and corneal reflexes absent	2 = incomprehensible sounds
0 = absent pupil, corneal, and cough reflex	1 = no verbal response
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 apnoea	

FOUR score = Full Outline of UnResponsiveness.

clinical decisions based on the scoring result. Thus, an inter-rater agreement within a range of ± 1 score points for both GCS and FOUR score was chosen as primary outcome. Secondary outcomes were exact inter-rater agreements and ratings of the sub-components of the two scores. For the primary outcome, a difference of 10% or more between the agreement rates of the neurologists and of the ICU staff was considered to be of clinical relevance. We estimated that scoring around 250 patients would allow to detect that difference with an α of 0.05 and a power of 90. Anticipating a drop-out rate of around 20% we planned to include 300 patients. We decided to analyse three pre-defined sub-groups for the primary endpoint: intubated patients, sedated patients, and patients with neurological diseases as primary admittance diagnosis. As previous work reported that the motor component of the GCS (GCS-mot) has a similar predictive value as the total GCS [5], and the combined eye and motor component of the FOUR score (FOUR-EM) has a similar predictive value as the total FOUR score [11] we separately analysed the predictive values for mortality and agreement rates for the GCS-mot and the FOUR-EM. Cronbach's α [12] was calculated to assess the internal consistency of both scores. Predictive values of the scores were assessed by calculating the area under the curve (AUC) with 95% confidence intervals from receiver operating characteristic (ROC) curves. Frequency tables were analysed using Fisher's exact test. A P less than 0.05 was considered to represent statistical significance.

Results

Patients

The study took place between May 2006 and April 2007. During the study period 992 patients were admitted to the subunit of our ICU where the study took place. In 664 cases, patients had to be excluded because no neurologist was available or patients were unwilling to participate. Scoring was performed on 328 patients. Of the 328, 61 (33 female; mean age 62 ± 17 years; APACHE II 13 ± 7) had to be excluded because no pair-wise rating occurred within a time interval of one hour. Thus, 267 patients (85 female; mean age 63 ± 17 years; APACHE II 14 ± 8) were included in the study resulting in 437 pair-wise ratings. Pair-wise ratings of the two neurologists were obtained in 100 of the 267 patients (40 female; mean age 64 ± 16 years; APACHE II 15 ± 7). The admittance diagnoses of the 267 included patients are displayed in Table 2. At the time of scoring 60 of 267 (22.5%) patients were intubated or had a tracheostoma and 52 of 267 (19.5%) received sedative drugs in the eight hours preceding scoring.

GCS vs. FOUR score

Overall 437 pair-wise ratings were analysed. Cronbach's α for the GCS (0.87) and the FOUR score (0.83) indicate a high degree of internal consistency for both scores. The frequency distribution of the GCS and FOUR scores are displayed in Figure 1. The agreement of the ratings in the three categories is displayed in Figure 2. Overall, there was a statistically significant difference (*P* = 0.0016) with regard to exact agreement between the GCS score (71%) and the FOUR score (82%) but not for the agreement within a range of ± 1 score point (GCS 90%; FOUR 92%). Tables 3 and 4 display the kappa values for the GCS and FOUR score, respectively. Note that the inter-rater agreement of the neurologists was significantly better than that of the ICU staff with regard to the FOUR score (Table 4) but not for the GCS (Table 3). No significant difference in inter-rater agreement was found for the three components of the GCS (Table 3). In the FOUR score, however, the inter-rater agreement significantly differed between the four components with the component 'respiration' achieving the highest agreement rates and the component 'brainstem' achieving the lowest agreement rates. In addition, the agreement between the neurologists for the components 'brainstem' and 'respiration' was significantly better than that between ICU staff (Table 4). Figure 3 displays the disagreement in pair-wise ratings for both scores. As a high proportion of scorings yielded maximum scores (Figure 1) and the agreement rates were highest at theses scores (Figure 3) we calculated kappa values after excluding the maximum scores (i.e. GCS 15 and FOUR score 16, respectively) and found a significant difference between the kappas with or without excluding the maximum scores for the GCS (kappa ± 95% confidence interval, 0.61 ± 0.05 vs. 0.48 ± 0.06) and the FOUR score (0.68 ± 0.05 vs. 0.54 ± 0.08).

Agreement between the neurologists

The two neurologists agreed exactly in 73% of the GCS scores and in 87% of the FOUR scores (*P* = 0.014). An agreement between the neurologists in the range of ± 1 point was observed for 88% of the GCS and for 91% of the FOUR scores, respectively (*P* = not significant (ns)). Cronbach's α showed a high internal consistency of the neurologists' ratings for both the GCS (α = 0.93) and the FOUR score (α = 0.88)

Table 2: Primary admittance diagnoses of 267 patients undergoing scoring of GCS and FOUR in intensive care

Reason for admission	N
Neurologic disorders	86
Cardiac disorders	74
Pulmonary disorders	33
Infectious diseases	33
Gastrointestinal disorders	15
Metabolic and endocrinologic disorders	7
Renal disease	1
Other	18

FOUR score = Full Outline of UnResponsiveness; GCS, Glasgow Coma Scale.

Agreement between neurologists and ICU staff

In 163 pair-wise ratings, ICU staff agreed exactly with the neurologist in 68% of the GCS scores (*P* = ns vs. agreement of the neurologists) and in 81% of the FOUR scores (*P* = 0.011 vs. GCS; *P* = 0.14 vs. agreement of neurologists). An agreement between the ICU staff and the neurologist in the range of ± 1 point was observed for 88% of the GCS and for 91% of the FOUR scores, respectively (*P* = ns for GCS vs. FOUR; *P* = ns vs. agreement of the neurologists).

Agreement between ICU staff

In 174 pair-wise ratings, ICU staff agreed exactly in 73% of the GCS scores (*P* = ns vs. agreement of the neurologists) and in 79% of the FOUR scores (*P* = 0.017 vs. GCS; *P* = 0.04 vs. FOUR score agreement of the neurologists). An agreement between ICU staff in the range of ± 1 point was observed for 93% of the GCS and for 88% of the FOUR scores, respectively (*P* = ns for GCS vs. FOUR; *P* = ns vs. agreement of the neurologists). The internal consistency of the ICU staffs' ratings was high for both the GCS (α = 0.87) and the FOUR score (α = 0.83).

Predictive value for 28-day mortality

Twenty eight-day mortality was 13%. There was no significant difference in the predictive values of the GCS (AUC of the ROC 0.78, 95% confidence interval 0.68 to 0.87), the FOUR score (AUC of the ROC 0.79, 95% confidence interval 0.69 to 0.89), and the APACHE II score (AUC of the ROC 0.86, 95% confidence interval 0.80 to 0.92) for 28-day mortality (Figure 4). However, mortality was significantly (*P* < 0.001) higher for patients with the three lowest total FOUR scores of 0 to 2 (83% died), when compared with patients with the lowest GCS score of 3 (45% died).

Analysis of predefined subgroups

The exact inter-rater agreement was better for the FOUR score than for the GCS in the three predefined subgroups intubated patients (n = 60; 78% vs. 65%, *P* = 0.026), sedated patients (n = 52; 73% vs. 62%, *P* = 0.095), and patients with neurological disease as primary admittance diagnosis (n = 86; 80% vs. 69%, *P* = 0.046). The exact inter-rater agreement of the neurologist and the ICU staff for the FOUR score was 79% vs. 68% for intubated patients, 88% vs. 74% for sedated patients, and 91% vs. 79% for patients with neurological disease as primary admittance diagnosis, respectively. Due to the compara-

Table 3: Weighted kappa values for the interrater agreement for the GCS

Rater pair	n	Total GCS	Eye response	Motor response	Verbal response
Neurologist-Neurologist	100	0.67 ± 0.10	0.75 ± 0.12	0.79 ± 0.10	0.78 ± 0.10
Neurologist-ICU staff	393	0.56 ± 0.09	0.68 ± 0.10	0.68 ± 0.10	0.70 ± 0.09
ICU staff- ICU staff	321	0.63 ± 0.08	0.74 ± 0.09	0.78 ± 0.09	0.86 ± 0.07
Overall	437	0.61 ± 0.05	0.72 ± 0.06	0.74 ± 0.06	0.78 ± 0.05

Data = weighted kappa ± 95% confidence interval. No statistically significant differences exist between rater pairs or the different components of the GCS.
GCS, Glasgow Coma Scale.

tively small absolute numbers these differences failed to reach statistical significance. There was no significant difference with regard to inter-rater agreement with a range of ± 1 point between the GCS and FOUR score or different kind of rater pairs for the predefined subgroups.

The AUC of the ROC of the GCS-mot for 28-day mortality was 0.75 (95% confidence interval 0.64 to 0.86) and did not significantly differ from the total GCS or the FOUR score. Over all 437 pair-wise ratings, the exact agreement for the GCS-mot was 87% ($P < 0.0001$ vs. the total GCS; $P = 0.006$ vs. FOUR score). The agreement within a range of ± 1 score point of the GCS-mot was 95% ($P = 0.0012$ vs. GCS; $P = 0.0002$ vs. FOUR score).

The AUC of the ROC of the combined FOUR-EM for 28-day mortality was 0.76 (95% confidence interval 0.66 to 0.87) and did not significantly differ from the total FOUR score, the GCS, or GCS-mot. Overall, 437 pair-wise ratings, the exact agreement for the FOUR-EM was 85% ($P = 0.07$ vs. total FOUR score; $P < 0.0001$ vs. GCS). The agreement within a range of ± 1 score point of the FOUR-EM was 92% ($P = 0.095$ vs. total FOUR score; $P = 0.21$ vs. GCS).

Discussion

The present study compared the inter-rater agreement of GCS and FOUR score as well as the inter-rater agreement of neurologist and ICU staff in unselected critically ill patients. In the primary outcome, i.e. the inter-rater agreement within the range of ± 1 score point, there was neither a significant difference between the GCS and the FOUR score nor a difference between neurologists and ICU staff. Exact inter-rater agreement was significantly better between neurologists than between ICU staff. Moreover, exact inter-rater agreement was significantly better for the FOUR score than for the GCS.

Recently, Wijdicks and colleagues, Wolf and colleagues, and Iyer and colleagues from the Mayo Clinic devised and validated the FOUR score [8,9,13]. Compared with the GCS, this new coma scale does not depend on a verbal response and provides greater neurological detail by inclusion of brainstem reflexes and breathing patterns. The present study is the first validation of the FOUR score in the ICU outside the institution that developed the FOUR score. In addition, the present study is the first validation of the FOUR score in unselected patients in a medical ICU. In agreement with the initial reports we

Table 4: Weighted kappa values for the interrater agreement for the FOUR score

Rater Pair	n	Total	Eye response	Motor response	Brainstem reflexes	Respiration
Neurologist-Neurologist	100	0.80 ± 0.09	0.85 ± 0.09	0.88 ± 0.09	0.87 ± 0.12	1.0 ± 0.00†
Neurologist-ICU staff	393	0.66 ± 0.09	0.77 ± 0.09	0.73 ± 0.09	0.71 ± 0.18	0.87 ± 0.08
ICU staff- ICU staff	321	0.63 ± 0.08*	0.85 ± 0.07	0.77 ± 0.09	0.53 ± 0.16*†	0.87 ± 0.08*
Overall	437	0.68 ± 0.05*	0.82 ± 0.05	0.78 ± 0.05	0.67 ± 0.10	0.90 ± 0.04

Data = weighted kappa ± 95% confidence interval. * $P < 0.05$ vs. neurologist-neurologist at same component of the score; † $P < 0.05$ vs. all other components of the FOUR score with the same raters.
FOUR score = Full Outline of UnResponsiveness.

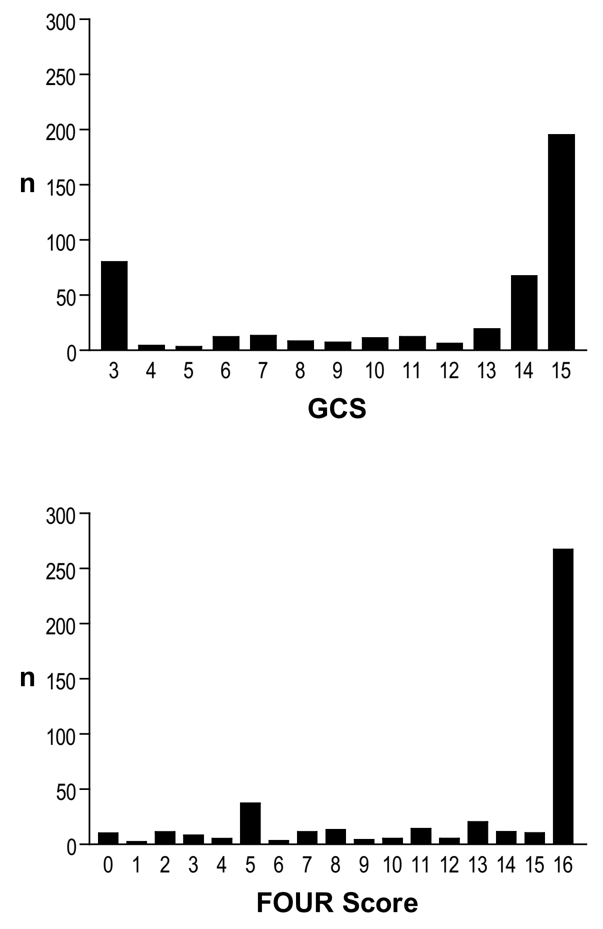

Figure 1 Frequency distribution of GCS and FOUR scores. Frequency distribution of 814 rated Glasgow Coma Scale (GCS) scores (top panel) and 814 rated Full Outline of UnResponsiveness (FOUR) scores (bottom panel).

observed that the inter-rater reliability for the FOUR score is at least as good as that of the GCS [8,9,13]. Moreover, our results demonstrate that the FOUR score is superior to the GCS with regard to exact inter-rater agreement. The inter-rater agreement in the present study was considerably lower for both GCS (kappa 0.59 vs. 0.82 to 0.98) and FOUR score (kappa 0.63 vs. 0.82 to 0.99) than previously reported [8,9,13]. This may be explained by the higher number of patients included in the present study, the inclusion of intubated and sedated patients, inherent variations in the level of consciousness among unselected ICU patients, organisational aspects of the scoring, and differences in the neurological expertise of the raters. Indeed, our neurologists achieved an inter-rater agreement for the FOUR score (kappa 0.80), but not for the GCS (kappa 0.67), comparable with that reported by Wijdicks and colleagues, Wolf and colleagues, and Iyer and colleagues [8,9,13]. Previous work demonstrated that the FOUR score predicts mortality as well as the GCS [8,9,11,13]. This is confirmed by our finding that the predictive value for 28-day mortality of the FOUR score equalled that of the GCS, and the APACHE II score. Moreover, in agreement with previous work our results demonstrate that mortality in medical ICU patients with the lowest FOUR score is higher than in patients with the lowest GCS.

The inter-rater agreement of the neurologists was never worse and partly significantly better than that of the ICU staff. However, as far as precision in scoring within the range of ± 1 score points is concerned, ICU staff equalled neurologists. This finding indicates that the precision in neurological scoring sufficient for the clinical settings achieved by general ICU staff cannot be significantly improved by dedicated specialists from outside the ICU.

The repetitive assessment of the level of consciousness is a routine procedure in ICU and so far the GCS is the most widely used tool. The present study confirms previous reports on a less than perfect inter-observer agreement of the GCS [2,14,15]. For the new FOUR score, the inter-rater agreement was never worse and partly better than that of the GCS. As the GCS is routinely performed in our unit, we were surprised and disappointed by the comparatively low inter-rater agreement of a longstanding standard procedure. To the best of our knowledge, there are no systematic data on the consistency of individual raters in repetitive ratings such as GCS in the ICU. Such a study would be very difficult to perform because within the time frame the level of consciousness could be kept reliably stable in critically ill patients most healthcare workers would not forget their previous scoring result. It is doubtful, however, that repetitive ratings are generally more precise than the pair-wise ratings reported in the present study. Thus, our findings suggest that in the clinical setting scores of individual patients should be cautiously interpreted taking into account both the dynamic course of critical illness and inter-rater and intra-rater disagreements.

Despite its limitations, the GCS has remained the standard coma scale over the past decades. In modern ICUs, multiple scores are repetitively used. Ideally, these scores should be simple, reliable, and predictive for relevant outcomes and/or relevant clinical decisions. With regard to these criteria, the present study revealed that the FOUR score is at least equivalent and partly even superior to the GCS. Given that the inter-rater reliability of the FOUR score between the neurologists was better than that between ICU staff and that the inter-rater reliability for the brainstem component of the FOUR score was significantly lower than for the other three components there is a potential for improvement for the inter-rater reliability of the FOUR score in the settings of the ICU. By contrast, our data reveal no such potential for improvement for the

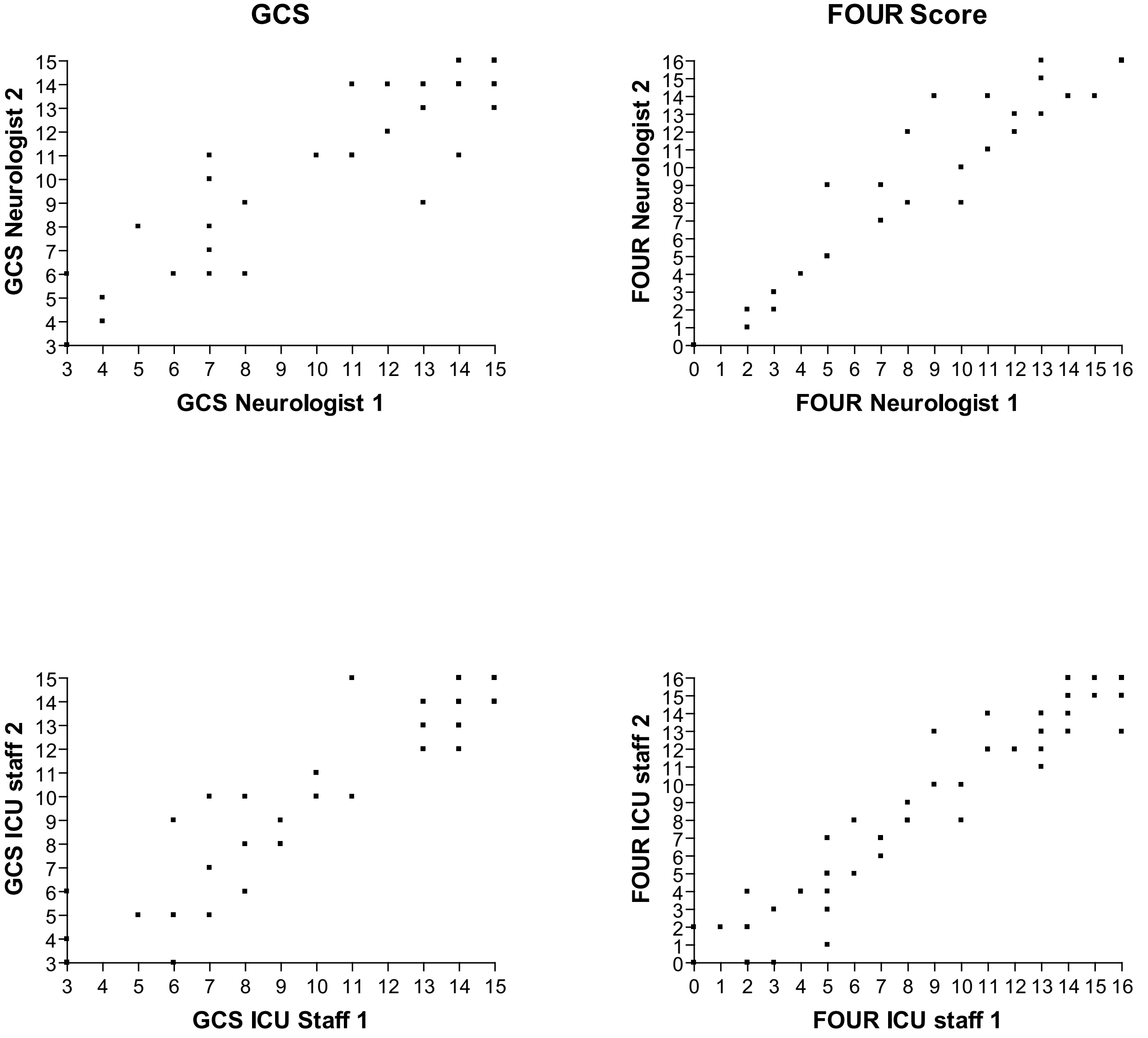

Figure 2 Inter-rater agreement of GCS and FOUR scores. Scatterplots of the pair-wise ratings of neurologists (top panels), ICU staff (middle panels), and neurologist-ICU staff (bottom panels) for the Glasgow Coma Scale (GCS; left side panels) and Full Outline of UnResponsiveness (FOUR) score (right side panels).

GCS. Compared with the GCS, the FOUR score contains more items. In addition, the brainstem categories of the FOUR score rely on up to three different items (pupil, corneal, and cough reflex) whereas all categories of the GCS rely on one item only. Thus, the FOUR score requires more time than the GCS and is more difficult to remember in acute situations. Although the FOUR score provides more neurological detail than the GCS it cannot replace a more in-depth neurological evaluation. Balancing the advantages and disadvantages of both scores, it is fair to state that the new FOUR score is a suitable alternative to the GCS. However, we think it is unlikely that the small advantage in inter-rater reliability will prompt intensivists to replace the GCS, a score with a long tradition in the ICU, by the new FOUR score.

Eken and colleagues reported, that in patients presenting with an altered level of consciousness, head trauma, or any neurological complaints on an emergency department, the FOUR-EM had a similar predictive value for unfavourable outcomes as the total FOUR score and the GCS [11]. Their finding is in agreement with the work of Gill and colleagues showing that the three individual GCS components alone performed similar to the total GCS score for the prediction of 4 clinically relevant TBI outcomes [5]. The present study confirms and extends these previous findings by demonstrating that among unse-

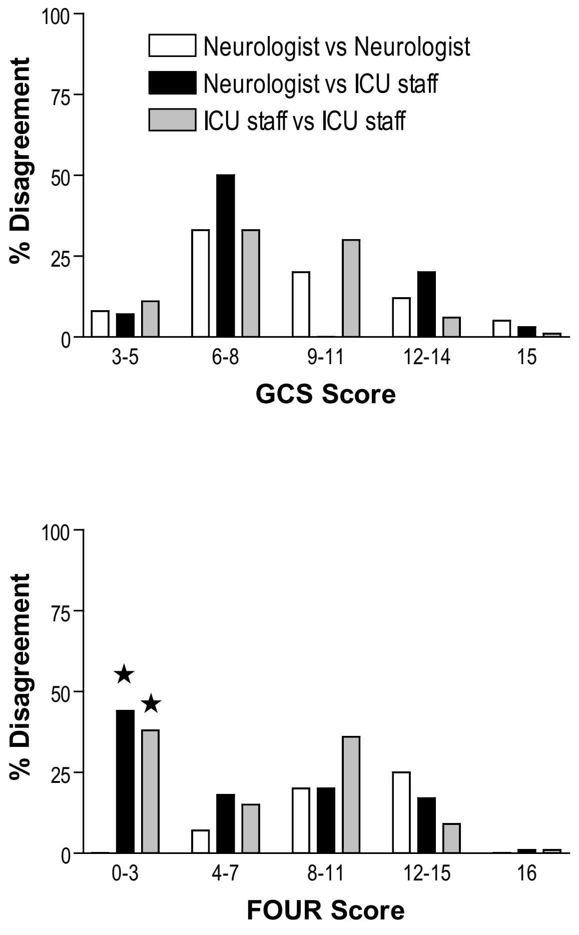

Figure 3 Disagreement rates for GCS and FOUR scores. Disagreements of more than one score point in pair-wise ratings of the Glasgow Coma Scale (GCS) score (top panel) and the Full Outline of UnResponsiveness (FOUR) score (bottom panel) respectively. Scores are divided into quartiles. As a substantial proportion of ratings were at the maximum of the each scale (i.e. GCS 15, FOUR 16), the maximum category is shown separately in addition to the quartiles. Disagreements are expressed as a percentage of the total number of ratings in a given quartile of the GCS score and FOUR score, respectively. White bars = disagreements between the neurologists; black bars = disagreements between the neurologists and ICU staff; grey bars = disagreements between ICU staff. For both scores, disagreements were significantly ($P < 0.001$) less frequent in the maximum category (i.e. GCS 15, FOUR 16) than in all other categories. * For the lowest quartile of the FOUR score, the disagreement between neurologist and ICU staff ($P = 0.034$) and between ICU staff and ICU staff ($P = 0.045$) was significantly greater than that between the neurologists.

lected critically ill patients in the medical ICU the predictive values of the FOUR-EM and of the motor component of the GCS for 28-day mortality does not differ from the total FOUR score or the GCS. Moreover, the inter-rater agreements for the FOUR-EM and the GCS-mot in the present study were better than for the total FOUR score and the GCS. Thus, reducing the complexity of a score can substantially improve inter-rater reliability without necessarily losing predictive power. In a multivariable analysis in over 8,000 head trauma patients Murray and colleagues [16] found that in addition to the GCS-mot, pupil reaction has an independent predictive value. Therefore, we tested whether adding the information on bilateral pupil reactivity to the FOUR-EM would significantly increase the predictive value for 28-day mortality, which was not the case.

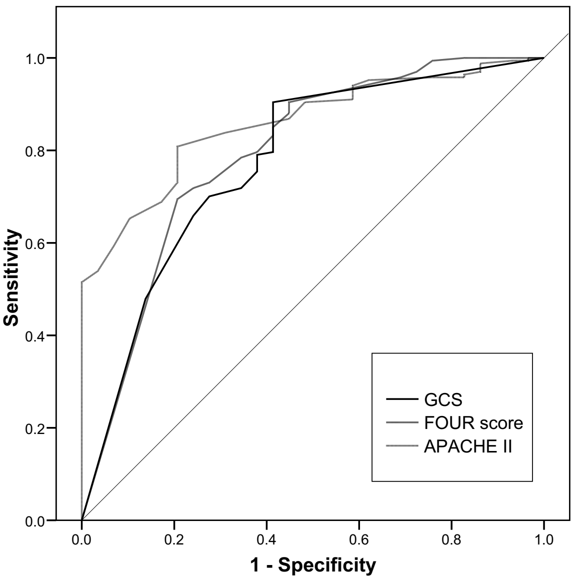

Figure 4 Predictive value for 28-day mortality. Receiver operating characteristic curve for the predictive value of Glasgow Coma Scale (GCS), Full Outline of UnResponsiveness (FOUR) score, and acute physiology and chronic health evaluation (APACHE) II score on 28-day mortality. There was no statistically significant difference between the areas under the curve of the three scores.

A limitation of our study is that due to the inclusion of unselected, and especially sedated, patients and the maximum time interval allowed for pair-wise ratings of one hour no perfectly stable experimental conditions for scoring were achieved. Particularly, we cannot exclude that some inter-rater disagreements are caused by true alterations in the level of consciousness. However, as raters performed the GCS and the FOUR score simultaneously such true alterations in the level of consciousness cannot explain the observed differences in the inter-rater agreement between the two scores. Moreover, our study conditions reflect the dynamic environment in the ICU so that our results give a fair estimate of the reliability of two coma scales in daily practice. A further limitation of our study is that no surgical patients, and especially no head trauma cases, were included so that the findings relate to unselected medical critically ill patients only. An inherent limitation of the validation of coma scales is the absence of an objective measure of the level of coma. Thus it should be kept in mind that better inter-rater reliability does not necessarily mean better accuracy.

Conclusions

The FOUR score performs better than the GCS with regard to exact inter-rater agreement, but not for the clinically more relevant agreement within the range of ± 1 score point or the predictive value for 28-day mortality. Although neurologists outperform ICU staff with regard to exact inter-rater agreement, the inter-rater agreement of ICU staff within the clinically more relevant range of ± 1 score point equals that of the neurologists. Thus, a precision in neurological scoring sufficient for the clinical settings cannot only be achieved by dedicated staff in specialised neuro-ICUs but also by ICU staff in general ICUs. The small advantage in inter-rater reliability of the FOUR score is most likely insufficient to replace the GCS, a score with a long tradition in the ICU.

Key messages

- The FOUR score, a new coma scale not relying on verbal response, performs better as the GCS with regard to exact inter-rater agreement, but not for the clinically more relevant agreement within the range of ± 1 score point.
- In neurological scoring, the inter-rater agreement within the range relevant for clinical decisions of ICU staff equals that of neurologists.

Abbreviations

APACHE: acute physiology and chronic health evaluation; AUC: area under the curve; FOUR: Full Outline of UnResponsiveness; FOUR-EM: combined eye and motor component of the FOUR score; GCS: Glascow Coma Scale; GCS-mot: motor component of the GCS; ROC: receiver operator characteristics.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

MF participated in the design of the study, performed the neurological scoring, and was responsible for data management. SR participated in the design of the study and performed the neurological scoring. AC performed the neurological scoring. MS was responsible for identifying suitable patients and performed the neurological scoring. AL participated in the design of the study and performed the neurological scoring. FT participated in the design of the study and performed the statistical analysis. PH participated in the design of the study and performed the neurological scoring. SM participated in the design of the study, performed the neurological scoring, and drafted the manuscript. All authors contributed to the interpretation of the results and read and approved the final manuscript.

Author Details

[1]Department of Medical Intensive Care, University Hospital, Spitalstrasse, Basel, 4031, Switzerland, [2]Department of Neurology, University Hospital, Spitalstrasse, Basel, 4031, Switzerland and [3]Department of Psychology, University of Neuchatel, Rue de la Maladière, Neuchatel, 2000, Switzerland

Received: 25 November 2009 Revised: 4 February 2010
Accepted: 14 April 2010 Published: 14 April 2010

References

1. Teasdale G, Jennett B: **Assessment of coma and impaired consciousness. A practical scale.** *Lancet* 1974, **2**:81-84.
2. Gill M, Martens K, Lynch EL, Salih A, Green SM: **Interrater reliability of 3 simplified neurologic scales applied to adults presenting to the emergency department with altered levels of consciousness.** *Ann Emerg Med* 2007, **49**:403-407.
3. Balestreri M, Czosnyka M, Chatfield DA, Steiner LA, Schmidt EA, Smielewski P, Matta B, Pickard JD: **Predictive value of Glasgow Coma Scale after brain trauma: change in trend over the past ten years.** *J Neurol Neurosurg Psychiatry* 2004, **75**:161-162.
4. Benzer A, Mitterschiffthaler G, Marosi M, Luef G, Puhringer F, De La Renotiere K, Lehner H, Schmutzhard E: **Prediction of non-survival after trauma: Innsbruck Coma Scale.** *Lancet* 1991, **338**:977-978.
5. Gill M, Windemuth R, Steele R, Green SM: **A comparison of the Glasgow Coma Scale score to simplified alternative scores for the prediction of traumatic brain injury outcomes.** *Ann Emerg Med* 2005, **45**:37-42.
6. Stanczak DE, White JG III, Gouview WD, Moehle KA, Daniel M, Novack T, Long CJ: **Assessment of level of consciousness following severe neurological insult. A comparison of the psychometric qualities of the Glasgow Coma Scale and the Comprehensive Level of Consciousness Scale.** *J Neurosurg* 1984, **60**:955-960.
7. Starmark JE, Stalhammar D, Holmgren E, Rosander B: **A comparison of the Glasgow Coma Scale and the Reaction Level Scale (RLS85).** *J Neurosurg* 1988, **69**:699-706.
8. Wijdicks EF, Bamlet WR, Maramattom BV, Manno EM, McClelland RL: **Validation of a new coma scale: The FOUR score.** *Ann Neurol* 2005, **58**:585-593.
9. Wolf CA, Wijdicks EF, Bamlet WR, McClelland RL: **Further Validation of the FOUR score coma scale by intensive care nurses.** *Mayo Clinic Proceedings* 2007, **82**:435-438.
10. Landis JR, Koch GG: **The measurement of observer agreement for categorical data.** *Biometrics* 1977, **33**:159-174.
11. Eken C, Kartal M, Bacanli A, Eray O: **Comparison of the Full Outline of Unresponsiveness Score Coma Scale and the Glasgow Coma Scale in an emergency setting population.** *Eur J Emerg Med* 2009, **16**:29-36.
12. Bland JM, Altman DG: **Cronbach's alpha.** *BMJ* 1997, **314**:572.
13. Iyer VN, Mandrekar JN, Danielson RD, Zubkov AY, Elmer JL, Wijdicks EF: **Validity of the FOUR score coma scale in the medical intensive care unit.** *Mayo Clin Proc* 2009, **84**:694-701.
14. Ely EW, Truman B, Shintani A, Thomason JW, Wheeler AP, Gordon S, Francis J, Speroff T, Gautam S, Margolin R, Sessler CN, Dittus RS, Bernard GR: **Monitoring sedation status over time in ICU patients: reliability and validity of the Richmond Agitation-Sedation Scale (RASS).** *JAMA* 2003, **289**:2983-2991.
15. Kho ME, McDonald E, Stratford PW, Cook DJ: **Interrater reliability of APACHE II scores for medical-surgical intensive care patients: a prospective blinded study.** *Am J Crit Care* 2007, **16**:378-383.
16. Murray GD, Butcher I, McHugh GS, Lu J, Mushkudiani NA, Maas AI, Marmarou A, Steyerberg EW: **Multivariable prognostic analysis in traumatic brain injury: results from the IMPACT study.** *J Neurotrauma* 2007, **24**:329-337.

doi: 10.1186/cc8963
Cite this article as: Fischer *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 *Critical Care* 2010, **14**:R64

Development of a modified paediatric coma scale in intensive care clinical practice

A Tatman, A Warren, A Williams, J E Powell, W Whitehouse

The Birmingham Children's Hospital NHS Trust: Department of Intensive Care
A Tatman
A Warren

Department of Paediatric Neurology
A Williams
W Whitehouse

Department of Public Health and Epidemiology, University of Birmingham
J E Powell

Correspondence to:
Dr W Whitehouse,
Department of Paediatric
Neurology, Birmingham
Children's Hospital,
Ladywood Middleway,
Birmingham B16 8ET.

Accepted 22 August 1997

Abstract

James' adaptation of the Glasgow coma scale (JGCS) was designed for young children. Intubated patients are not allocated a verbal score, however, so important changes in a patient's conscious level may be missed. A grimace score was therefore developed and assessed for use in intubated children.

Two observers made a JGCS observation within 15 minutes of each other. One observer was the patient's nurse and the other a trained investigator. Interobserver reliability was determined between the first and second observation for each component of the scale. Reliability was measured using κ and weighted κ statistics.

Seventy three children had 104 sets of observations. Interobserver reliability was moderate to good for all components, with the grimace score better than the verbal score.

It is concluded that the grimace score is more reliable than the verbal score and may be useful in intubated patients in whom the verbal score cannot be used.

(*Arch Dis Child* 1997;**77**:519–521)

Keywords: coma scale; intensive care; conscious level

The Glasgow coma scale has been widely adopted in the management of adult and paediatric coma.[1, 2] It should not be used in small children as the verbal component is not appropriate.[3] Several coma scores have been developed specifically for children in an attempt to compensate for their differences in verbal and motor capabilities.[4–11] Three years ago, we introduced into our intensive care unit (ICU) a modified Glasgow coma scale, which is Sharples' adaptation (personal communication) of the James' adaptation of the Glasgow coma scale (JGCS) (table1).[12]

During this time, our nursing staff reported that many children who were intubated showed varying degrees of orofacial grimacing when stimulated. Therefore we developed a grimace score to replace the verbal component in intubated children. We report the results of a study to assess the reliability of our modified coma scale in this clinical setting.

Subjects and methods

STUDY DESIGN

After receiving local ethical committee approval, children on the ICU with coma from any cause were selected in a quasirandom manner: whenever one of the three trained investigators was available, the patient accessible, and the patient had not been studied within 24 hours nor with the same JGCS (on the routine nursing JGCS chart).

Verbal consent was obtained from parents when available. A set of observations consisted of two JGCS (table 1) scores, the second score being completed within 15 minutes of the first. These were performed sequentially by two observers, one being the child's bedside nurse

Table 1 Modified Glasgow coma scale. Pain as nail bed pressure with pencil; score best response

Adult and child > 5 years	*Child < 5 years*
Eye opening	
E4 spontaneous	As older child
E3 to verbal stimulus	As older child
E2 to pain	As older child
E1 no response to pain	As older child
Verbal	
V5 orientated	Alert, babbles, coos, words or sentences to usual ability
V4 confused	Less than usual ability or spontaneous irritable cry
V3 inappropriate words	Cries to pain
V2 incomprehensible sounds	Moans to pain
V1 no response to pain	No reponse to pain
VT intubated	Intubated
Grimace	
G5 spontaneous normal facial/oromotor activity, for example sucks tube, coughs	
G4 less than usual spontaneous ability or only responds to touch	
G3 vigorous grimace to pain	
G2 mild grimace or some change in facial expression to pain	
G1 no response to pain	
Motor	
M6 obeys commands	Normal spontaneous movements or withdraws to touch
M5 localises to pain stimulus	As older child
M4 withdraws from pain	As older child
M3 abnormal flexion to pain	As older child
M2 abnormal extension to pain	As older child
M1 no response to pain	As older child

and the other being one of three trained observers. The observers were blinded to the preceding score. Children who were not intubated were given a verbal score. Children who were intubated were given a grimace score. We excluded children with cervical spinal cord injury, peripheral nerve disease, or neuromuscular disorders, including residual paralysis from neuromuscular blockade. The painful stimulus was nail bed pressure on both upper limbs, using a pencil. The best response was taken for the observation.

Table 2 Diagnostic category on admission to ICU

Diagnosis	*Number*
Cardiac surgery	30
General surgery	12
Neurosurgery	5
Metabolic	7
General medical	3
Neurology	10
Head injury	6

Table 3 Each pair of observations for each component of the adapted JGCS

Eye opening E1–E2

	E2 score			
	1	*2*	*3*	*4*
E1 score				
1	35	6	1	1
2	5	8	0	5
3	0	2	0	3
4	2	5	3	24

Verbal V1–V2

	V2 score				
	1	*2*	*3*	*4*	*5*
V1 score					
1	5	1	0	0	0
2	0	1	1	1	0
3	0	0	1	1	0
4	2	1	1	3	1
5	1	0	3	0	5

Grimace G1–G2

	G2 score				
	1	*2*	*3*	*4*	*5*
G1 score					
1	14	2	0	1	0
2	6	9	4	1	1
3	0	2	5	3	1
4	0	2	3	1	0
5	0	2	0	0	11

Motor M1–M2

	M2 score					
	1	*2*	*3*	*4*	*5*	*6*
M1 score						
1	5	0	0	3	0	0
2	0	0	1	2	1	0
3	0	1	2	4	0	0
4	3	1	2	24	10	2
5	1	0	0	6	12	4
6	0	0	0	2	8	10

Table 4 Each pair of observations for the summated adapted JGCS, with grimace in place of verbal

Summated EGM1–EGM2

	EGM1 score												
	3	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
EGM2 score													
3	3	0	0	1	0	0	0	0	0	0	0	0	0
4	0	0	0	2	0	0	0	0	0	0	0	0	0
5	0	0	1	1	0	0	0	0	0	0	0	0	0
6	0	0	1	2	2	1	0	1	0	0	0	0	0
7	0	1	0	3	2	0	0	0	0	0	0	0	0
8	1	0	0	0	3	2	2	0	1	0	0	0	0
9	0	0	0	1	0	2	3	2	0	0	0	0	0
10	0	0	0	0	1	2	2	1	1	0	1	0	0
11	0	0	0	0	0	0	1	0	1	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	2	0	0
13	0	0	0	0	0	0	0	0	0	2	0	0	2
14	0	0	0	0	0	0	0	1	0	0	1	2	2
15	0	0	0	0	0	0	0	0	1	0	0	2	3

Table 5 Interobserver agreement

	No	*κ (95% CI)*	*Weighted κ (95% CI)*
E1–E2	100	0.50 (0.38 to 0.63)	0.64 (0.53 to 0.76)
V1–V2	28	0.41 (0.20 to 0.63)	0.49 (0.25 to 0.73)
G1–G2	68	0.50 (0.32 to 0.61)	0.63 (0.50 to 0.77)
M1–M2	104	0.33 (0.20 to 0.46)	0.49 (0.36 to 0.62)
EVM1–EVM2	28	0.29 (0.10 to 0.48)	0.57 (0.39 to 0.75)
EGM1–EGM2	68	0.25 (0.13 to 0.38)	0.69 (0.60 to 0.78)

CI=confidence interval.

STATISTICAL ANALYSIS

Interobserver reliability (E1–E2, V1–V2, G1–G2, M1–M2, and summated scores EVM1–EVM2 and EGM1–EGM2), that is, the level of agreement between the two observations, was measured by the κ and weighted κ statistics.[13] While the κ statistic measures the *level* of agreement above that expected by chance, it does not take into account the *degree* of disagreement between observations. The weighted κ statistic measures agreement and takes into account the magnitude of the disagreement.

For both κ and weighted κ, strength of agreement is interpreted as < 0.2 = poor; 0.21–0.40 = fair; 0.41–0.6 = moderate; 0.61–0.80 = good; > 0.8 = very good or near perfect.

Results

One hundred and four sets of observations were completed in 73 children of whom 42 were boys. Four children had severe orbital swelling and were not given an eye score. Forty one observers were involved (38 nurses and three trained observers). The children ranged in age from 1 day to 16 years (median age 73 days). Table 2 shows the diagnostic categories. Tables 3 and 4 show the raw data for each component and for summated scores using the grimace and verbal scores separately. Table 5 shows the interobserver reliability.

Discussion

We adopted the JGCS because it takes account of developmental immaturity in small children, uses the same number of points irrespective of the child's age, and is simple for the patient's nurse to use without additional staff or equipment.

Several studies have examined the reliability of paediatric coma scales using two or three trained observers.[5 10 14] This is useful for determining a scale's experimental reliability, but may not necessarily translate into clinical practice.[15] For example, in our ICU there are over 100 nurses with varying levels of experience. Therefore, any scale must be robust enough to produce reliable results given the observers who will be using it. Complicated scales, which are used relatively infrequently, are unlikely to be reliable.

Our results suggest that despite a large number of observers, there is moderate to good interobserver agreement for the components of this scale.

The grimace component appears to be more reliable than the verbal component. They may measure different aspects of brain function and cannot necessarily be equated clinically. Facial expression, however, is an important part of non-verbal communication, so facial grimace and verbal language are not totally independent skills. Furthermore, we believe that in intubated patients the restoration of a third variable (eye opening, motor, *and grimace*) in assessing coma increases the likelihood of detecting an improvement or a deterioration in the patient's condition, particularly when the variables are measured independently.

We have included the summated values for interest. We do not summate the values clinically, as the variables have different weights and are not clinically comparable.[16 17]

Although the grimace score has not been validated for outcome, it is more reliable than the verbal score in this study and may be useful in intubated patients when the verbal score cannot be used.

This study has also shown the reliability of the other components of our adaptation of the JGCS when used by nurses and doctors in an ICU.

We are very grateful for the help of all the nurses at Birmingham Children's Hospital ICU for their contribution to the study and their continuing support.

1 Teasdale G, Jennett B. Assessment of coma and impaired consciousness: a practical scale. *Lancet* 1974;**ii**:81-4.

2 Jennett B, Teasdale G. Aspects of coma after severe head injury. *Lancet* 1977;**i**:878-81.

3 Raimondi AJ. Editorial comment. *Childs Nerv Syst* 1988;**4**: 39-40.

4 Reilly PL, Simpson DA, Sprod R, Thomas L. Assessing the conscious level in infants and young children: a paediatric version of the Glasgow coma scale. *Childs Nerv Syst* 1988; **4**:30-3.

5 Simpson DA, Cockington RA, Hanieh A, Raftos J, Reilly PL. Head injuries in infants and young children: the value of the paediatric coma scale. *Childs Nerv Syst* 1991;7:183-90.

6 Hahn YS, Chyung C, Barthel MJ, Bailes J, Flannery AM, McLone DG. Head injuries in children under 36 months of age—demography and outcome. *Childs Nerv Syst* 1988;**4**: 34-40.

7 Raimondi AJ, Hirschauer J. Head injury in infants and toddlers. *Child Brain* 1984;**11**:12-35.

8 Gordon NS, Fois A, Jacobi G, Minns RA, Seshia SS. The management of the comatose child-consensus statement. *Neuropediatrics* 1983;**14**:3-5.

9 Morray JP, Tyler DC, Jones TK, Stuntz JT, Lemire RJ. Coma scale for use in brain-injured children. *Crit Care Med* 1984;**12**:1018-20.

10 Yager JY, Johnston B, Seshia SS. Coma scales in pediatric practice. *Am J Dis Child* 1990;**144**:1088-91.

11 Lovejoy FH Jr, Smith AL, Bresnan MJ, Wood JN, Victor DI, Adams PC. Clinical staging in Reye syndrome. *Am J Dis Child* 1974;**128**:36-41.

12 James HE, Trauner DA. The Glasgow coma scale. In: James HE, Anas NG, Perkin RM, eds. *Brain insults in infants and children*. Orlando: Grune and Stratton, 1985:179-82.

13 Altman DG. Some common problems in medical research 14.3 inter-rater agreement. In: Altman DG, ed. *Practical statistics for medical research*. London: Chapman and Hall, 1991:403-8.

14 Newton CRJ, Kirkham FJ, Johnston B, Marsh K. Interobserver agreement of the assessment of coma scales and brainstem signs in non-traumatic coma. *Dev Med Child Neurol* 1995;**37**:807-13.

15 Rowley G, Fielding K. Reliability and accuracy of the Glasgow coma scale with experienced and inexperienced users. *Lancet* 1991;**337**:535-8.

16 Teasdale G, Jennett B, Murray L, Murray G. Glasgow coma scale: to sum or not to sum. *Lancet* 1983;**ii**:678.

17 Jagger J, Jane JA, Rimel R. The Glasgow coma scale: to sum or not to sum. *Lancet* 1983;**ii**:97.

:tal and placental size
990; 301: 259–63.
ci J. Socioeconomic
iorbidity caused by
d. *Soc Sci Med* 1985;

; in death—specific
1003–06.
rwegian experience.

self perceived health
pidemiol Community

AG. Migration and
c in Great Britain.

.G. Migration and
Br Med J 1990; 300:

. *J Prevent Soc Med*

il unemployment in
man Health 1980; 34:

iemic heart disease.

ude and causes of
evidence from the
990; 44: 265–70.
s in childhood and
Med J 1990; 301:

ns in childhood and
J 1991; 302: 113–14.
lker M. Social class
en. *Lancet* 1987; ii:

ontrol of growth in

: Britain. OPCS SS

From cardiovascular
of birth in England
4: 139–41.

Jnited Kingdom
medical science;
c; but when their
ie prosecution of
ive fallen, within
, are taken into
it the discoveries
e collected, than
hospitals about
; and it has been
in the United
e the number to
ases have been
: hospital books?
ficers, whom Sir
titioners of the
ice, their fidelity
ty? What would
plan of hospital
duty than the
ed that medicine
Histories of cases
, and those cases
es it happen that
e applications, a
of superstition;

(Jan 30, 1841)

CLINICAL PRACTICE

Reliability and accuracy of the Glasgow Coma Scale with experienced and inexperienced users

GLENN ROWLEY KATY FIELDING

To investigate whether the Glasgow Coma Scale (GCS) can be used reliably and accurately by inexperienced observers, ratings made by observers grouped by level of experience were examined for within-group interobserver disagreements and for discrepancies with scores given by an expert. The GCS was used accurately by experienced and highly trained users, but inexperienced users made consistent errors. The errors were such that they would not be detectable by studies that examine only interobserver agreement, and they were substantial, averaging in some cases more than one point on the four-point and five-point scales of the GCS. Also, the error rates were highest at the intermediate levels of consciousness, for which the detection of changes in condition is vital. The findings support the continued use of the GCS by appropriately qualified personnel, but call into question much of the conventional wisdom about its reliability when used by untrained or inexperienced staff. The findings also suggest that interobserver comparisons are insufficient for establishing the viability of the GCS.

Lancet 1991; **337**: 535–38.

Introduction

Since its introduction in 1974 the Glasgow Coma Scale (GCS)[1] has gained widespread acceptance around the world as a means of assessing the level of consciousness of patients with head injury. The scale consists of ratings for eye opening based a four-point scale and those of verbal response and motor response on five-point scales. Its primary purpose is to alert medical and nursing staff to deterioration in a patient's neurological status. The report by Teasdale, Knill-Jones, and Van der Sande[2] has widely been accepted as evidence that the scale is reliable when used by experienced or inexperienced people. Teasdale's assertion[3] that the GCS "has proved to be useful, reliable and practical and can be used by personnel of all grades of experience" has been widely endorsed,[4-7] even though the study had revealed disagreement rates as high as 0·191,[2] which indicates average disagreements among observers approaching a whole scale point. Subsequent research on the GCS has, with few exceptions,[8,9] taken the reliability of the scale as assured.

The establishment of a high level of observer agreement is a necessary but not sufficient condition for continued faith in the GCS. This study addresses questions that include, but go beyond, conventional questions of reliability. First, it looks at disagreements between observers as just one of several sources of variation in GCS ratings, and second, it looks at the accuracy of ratings, by comparing the ratings of observers with those made by one expert observer, and determining where and under what circumstances important errors occur.

Methods

Study design

The observers were four groups of nurses. Group 1 consisted of 3 experienced nurses with at least 2 years' post-registration experience and at least 1 year of current neuroscience nursing practice. 2 held certificates in neuroscience nursing, and all had undertaken formal instruction in the GCS before this study. In a previous study[10] this group had shown extremely high levels of inter-rater agreement in using the GCS. Group 2 consisted of 7 newly graduated nurses; all had received instruction in the use of the GCS, but had had very little experience with it and had not previously worked in a neurosurgical ward. Groups 3 and 4 (5 and 6 student nurses, respectively) had not previously worked in a neurosurgical ward, nor had they received specific instruction in neurological assessment.

Accompanied but not assisted by the expert rater (an experienced charge nurse responsible for instruction in the use of the GCS in the hospital), every group used the GCS to assess the same five or six patients on three to five occasions. Different patients were used for different groups of nurses to protect patients from over-assessment. All patients had undergone neurosurgery or had sustained cranial trauma and were being cared for in a neurosurgical ward. They were selected for the study simply because they were available for assessment at the time that a group of observers could be brought together.

Before data collection, the group of observers was given 10 minutes to read instructions on the correct use of the GCS. These instructions set out the protocol to be followed, specifying the stimuli to be used and the order in which observations were to be taken. The expert observed all procedures and made written notes on any departures from protocol as they occurred; then when the group had dispersed, she made her own GCS observations.

Data from any patient not available for all occasions were not included in the study. The amount of unusable data was small. Groups 1 and 3 were unaffected; 1 patient was unavailable on the last day in group 2, and 1 patient was unavailable on the last 3 days and 1 on a single day in group 4. Complete data were obtained as follows: group 1—6 patients observed by 3 raters on 4 occasions; group 2—5 patients, 7 raters, 3 occasions; group 3—5 patients, 5 raters, 5 occasions; and group 4—5 patients, 6 raters, 5 occasions. GCS ratings for eye response and motor response were recorded separately for the left and right sides, in accord with the usual practice at the hospital. Consequently, each patient assessment yielded five ratings: left and right eye response, verbal response, and left and right motor response. Separate analyses were conducted for each of these five aspects, and for each of the four groups of raters.

ADDRESSES: School of Graduate Studies, Faculty of Education, Monash University, Melbourne, Victoria 3168, Australia (Glenn Rowley, PhD); Neurological/Neurosurgical Unit, Alfred Hospital, Melbourne (Katy Fielding, RN). Correspondence to Dr G. Rowley.

[redacted]

Measurement of observer agreement

The three measures of observer agreement used were: percent agreement, a reliability coefficient, and the "disagreement rate" introduced by Teasdale et al.[2]

Percent agreement is widely understood but it is an "all-or-nothing" measure and is not sensitive to the amount of variability among the patients being observed or to the magnitudes of the errors made.

A *reliability coefficient* expresses the variance associated with real differences in that being measured as a proportion of the total variance in a set of data. The coefficients reported in this study are derived from generalisability theory,[11,12] which extends and broadens standard reliability theory by acknowledging that all measurements are subject to multiple sources of error and, with adequate designs (as in this study), enables their contributions to measurement error to be quantified. The method of computation is as outlined by Shavelson et al;[12] for these data all variation associated with patients or their interaction with occasions is treated as systematic variance, and all variation associated with observers or their statistical interactions with other sources, as error. These coefficients are sensitive to the magnitudes of any disagreements, and assess them in relation to the amount of variation present among the patients being graded. The resulting coefficient can range anywhere from zero (no systematic variance) to one (no error variance).

The *disagreement rate* as calculated by Teasdale et al's formula,[2] from discrepancies between individual ratings and "consensus" ratings (ie, the modal observations for that patient), is sensitive to the magnitudes of the disagreements between observers, not only to the number of disagreements. However, it is an ad hoc measure created for a specific purpose, and it is not widely known or understood by researchers in other areas. It indicates the average discrepancy between a single observation on a patient and the mode of the observations by all observers on that patient, expressed as a fraction of the maximum possible distance from the modal (consensus) rating.

Measurement of accuracy

Accuracy was addressed by comparing the ratings given by each rater to those awarded by the expert rater. The ratings of the expert observer were taken as "correct" (ie, the best available approximation to the truth), a procedure in line with those traditionally used to establish the validity of psychological measurements,[13] but rarely in the assessment of accuracy of GCS.[14]

Computation of disagreement rate from the discrepancies between individual ratings and the ratings of the expert observer was identical to that used for the Teasdale disagreement rate, except that the expert's rating is substituted for the consensus rating; it indicates degree of disagreement between raters and the expert observer, and a low value indicates that the ratings are accurate. Directional and absolute means of discrepancies between raters' and expert's scores are presented. For the directional means, a positive discrepancy occurs when the rating given is too high, and a negative discrepancy when it is too low.

Identification of sources of error

Since patients who are fully alert (rating 4 on eye response, 5 on verbal and motor response) or profoundly unconscious (rating 1) can be rated much more easily than can patients whose conditions lie between these two extremes, data were also analysed by category—end-of-range (4 and 1 for eye response, 5 and 1 for verbal and motor response) and middle-range.

Results

Within-group reliability

Group 1 (the experienced nurses) were the most consistent in their assessments, but only marginally more so than were groups 2 (new graduates) and 3 (student nurses) (table I). Overall group 4 showed lower levels of agreement than did the other three groups because of disagreement on the rating of eye opening; for verbal response, and for left and right motor response, their agreement was similar to those within the other groups.

TABLE I—DISAGREEMENT RATES, PERCENT AGREEMENT, AND RELIABILITY COEFFICIENTS FOR FOUR GROUPS OF OBSERVERS

Measure of agreement with with other raters	Group 1 (experienced nurses)	Group 2 (new graduates)	Group 3 (student nurses)	Group 4 (student nurses)
Percent agreement				
Left eye	98·6	96·2	92·5	82·0
Right eye	98·6	96·2	92·5	77·3
Verbal	100·0	95·2	89·2	92·7
Left motor	98·4	94·3	99·2	92·7
Right motor	94·4	93·3	100·0	94·0
(Mean)	(98·0)	(95·0)	(94·7)	(88·5)
Reliability coefficient				
Left eye	0·994	0·983	0·789	0·757
Right eye	0·994	0·983	0·789	0·757
Verbal	1·000	0·984	0·937	0·953
Left motor	0·996	0·906	0·993	0·879
Right motor	0·985	0·872	1·000	0·955
(Mean)	(0·944)	(0·946)	(0·902)	(0·860)
Disagreement rate				
Left eye	0·005	0·016	0·061	0·110
Right eye	0·005	0·016	0·106	0·163
Verbal	0·000	0·025	0·035	0·023
Left motor	0·007	0·048	0·004	0·047
Right motor	0·030	0·064	0·000	0·028
(Mean)	(0·009)	(0·034)	(0·032)	(0·074)

Accuracy

The disagreement rate and percent agreement computed around the ratings of the expert observer warrant concern (table II). Although group 1 (the experienced nurses) maintained a high level of accuracy (their disagreement rates averaged 0·026 and their percent agreements averaged 96·4%), the other three groups gave observer agreements of only 60–80% and disagreement rates of approximately 0·15 to 0·30.

Two conclusions follow from our findings. First, experienced and well-trained practitioners can use the GCS with extremely high levels of reliability and accuracy. Second, practitioners with limited training and experience in the use of the instrument can use it with high levels of reliability, but the accuracy of their ratings is suspect.

Magnitude and source of errors

On all scales, the mean discrepancies for group 1 were close to zero, whereas those for groups 2, 3, and 4 were

TABLE II—DISAGREEMENT RATES AND PERCENT AGREEMENT COMPUTED AROUND THE EXPERT RATING FOR FOUR GROUPS OF OBSERVERS

Measure of agreement with expert	Group 1	Group 2	Group 3	Group 4
Percent agreement				
Left eye	94·4	84·8	83·3	72·7
Right eye	94·4	84·8	83·3	66·7
Verbal	100·0	81·9	70·8	92·0
Left motor	98·6	67·6	59·2	38·0
Right motor	94·4	71·4	58·3	38·7
(Mean)	(96·4)	(78·1)	(71·0)	(61·6)
Disagreement rate				
Left eye	0·046	0·108	0·101	0·154
Right eye	0·046	0·108	0·101	0·242
Verbal	0·000	0·079	0·188	0·049
Left motor	0·007	0·248	0·408	0·520
Right motor	0·032	0·237	0·389	0·577
(Mean)	(0·026)	(0·156)	(0·237)	(0·308)

TABLE III—MEAN DISCREPANCIES BETWEEN RATERS AND EXPERT JUDGES

Mean discrepancy with expert rater	Group 1	Group 2	Group 3	Group 4
Directional				
Left eye	0·111	−0·286	−0·144	−0·133
Right eye	0·111	−0·286	−0·144	0·133
Verbal	0·000	0·105	0·552	0·147
Left motor	−0·014	−0·400	−0·784	−0·933
Right motor	−0·014	−0·590	−0·600	−1·207
(Mean)	(0·039)	(−0·291)	(−0·224)	(−0·399)
Absolute				
Left eye	0·139	0·286	0·240	0·360
Right eye	0·139	0·286	0·240	0·653
Verbal	0·000	0·181	0·712	0·147
Left motor	0·014	0·552	0·784	1·080
Right motor	0·042	0·667	0·760	1·233
(Mean)	(0·067)	(0·394)	(0·547)	(0·695)

TABLE IV—PERCENT AGREEMENT WITH EXPERT RATINGS FOR END-OF-RANGE AND MIDDLE-OF-RANGE SCORES

Level of consciousness*	Group 1	Group 2	Group 3	Group 4	All groups
Left eye					
End	93·0 *(57)*	90·5 *(84)*	91·8 *(110)*	84·3 *(102)*	89·5 *(353)*
Middle	100·0 *(15)*	61·9 *(21)*	26·7 *(15)*	47·9 *(48)*	55·5 *(99)*
Right eye					
End	93·0 *(57)*	90·5 *(84)*	91·8 *(110)*	73·3 *(126)*	88·1 *(377)*
Middle	100·0 *(15)*	61·9 *(21)*	26·7 *(15)*	33·3 *(24)*	55·5 *(75)*
Verbal					
End	100·0 *(66)*	98·6 *(70)*	71·2 *(125)*	95·8 *(144)*	89·4 *(405)*
Middle	100·0 *(6)*	48·6 *(35)*	.. *(0)*	0·0 *(6)*	74·2 *(47)*
Left motor					
End	100·0 *(63)*	94·3 *(70)*	100·0 *(75)*	100·0 *(54)*	98·5 *(262)*
Middle	88·9 *(9)*	14·3 *(35)*	2·0 *(50)*	3·1 *(96)*	9·5 *(190)*
Right Motor					
End	100·0 *(45)*	64·8 *(91)*	100·0 *(75)*	90·0 *(60)*	86·0 *(271)*
Middle	96·3 *(27)*	85·7 *(14)*	2·0 *(50)*	3·3 *(90)*	23·2 *(181)*
Combined					
End	97·2 *(288)*	86·7 *(399)*	89·1 *(495)*	87·2 *(486)*	89·4 *(1668)*
Middle	97·2 *(72)*	47·6 *(126)*	7·7 *(130)*	14·3 *(264)*	30·2 *(592)*

*End = end of range of ratings; middle = middle of range.
Numbers in parentheses refer to the numbers of observations obtained within each category.

substantially larger and usually negative (table III)—ie, the ratings given by the less experienced groups were lower than they should have been, which indicates that their failure lay in not noticing, or not producing, a response that was detected by the experienced practitioners. Except for group 1, all groups made substantial errors, and in group 4, the size of the error in rating motor response averaged more than one scale point, for both sides. Motor response clearly caused the greatest problem for the inexperienced raters; groups 2 and 3 were only marginally better than group 4, being in error, on average, by half to three-quarters of a scale point. The experienced users were clearly superior on all aspects involved, and their mean errors were all small fractions of a scale point.

There were 1668 observations in the end-of-range category, and 592 in the middle-range category. The high level of accuracy attained by the experienced raters was maintained whether the ratings were easy (end-of-range) or more difficult (middle-of-range) (table IV). The other three groups rated accurately for the end-of-range scores, but much less so for the middle-range scores; in some cases the level of agreement was even less than could be expected to occur by chance.

Discussion

The findings in this study provide the strongest support yet seen for the use of the GCS by experienced and highly trained observers. The disagreement rates were very low, compared with, for example, those found by Teasdale et al,[2] who reported disagreement rates of 0·089 for eye opening, 0·091 for verbal response, and 0·091 for motor response (in their paper no distinction was made between left and right). Teasdale's observers (6 nurses, 7 neurosurgeons, and 5 general surgical trainees) had not been trained to use the GCS but were provided with standard definitions as guidance. Our nurse-observers ranged from thoroughly trained and experienced nurses to inexperienced nursing students with minimal formal training, so the high levels of agreement are perhaps surprising but consistent with the oft-made[3-6] statement that the GCS allows accurate assessment by both experienced and inexperienced staff.

Our findings also illustrate an observation noted by Fielding and Rowley[10]—that disagreement rates create a more favourable impression than does percent agreement. The reason is that, in computing a disagreement rate, discrepancy is counted as a fraction of the greater of the distances from the modal rating to either end of the scale, whereas in computing a percent agreement, each discrepancy is counted as one. The two are, therefore, not directly comparable. A percent agreement of, say, 95·0% represents a considerably higher level of agreement than does a disagreement rate of 0·05.

However, the results obtained for inexperienced observers call into question much of the previous research and conventional wisdom on the reliability of the GCS when it is used by inexperienced staff. We found impressive levels of consistency among inexperienced observers, but there was clear evidence of disagreement between their ratings and those of an expert observer.

Agreement with the expert observer is a much tougher test of the GCS than is interobserver agreement, and the results reflect this difference. In this study, as in others, measures of interobserver agreement were derived from observations of the same patients' responses to the same stimuli, whether applied correctly or incorrectly. Measures of accuracy were derived from comparisons between these observations and those of an expert observer to stimuli applied (we assume correctly) by the expert. Consequently any lack of expertise in application of the stimuli would contribute to error in the assessments of accuracy, but could not be detected by measures of interobserver agreement.

The results demonstrate that the inexperienced users, although maintaining a high level of agreement amongst themselves, made substantial and serious errors, averaging up to one point on the four-point and five-point scales of the GCS. Only studies that use an expert observer, such as this one and another by Ingersoll and Leyden,[14] can identify common errors among a set of observers. The greatest difficulty that inexperienced users had was with middle-of-range scores, which confirms the suggestion made by Starmark, Holmgren, and Stalhammer.[15] The assertion that the scale is usable by personnel of all grades of experience has never been supported by evidence of accuracy, and our results do not support it.

The method adopted in this study is applicable to category rating scales in general, and has wide application. Generally, our results indicate that typical observer agreement studies are essentially uninterpretable as they stand. For the GCS in particular, we found that measures of

[redacted]

disagreement were highly dependent on the conditions of the patients observed. Comparisons between measurements of disagreement obtained from different patients may not be valid if it is likely that their conditions differ substantially. In the present study, the tasks presented to the four groups were not of equal difficulty, with group 4 having a smaller percentage of end-of-range observations to make (65%) than the other three groups (76 to 80%). As a minimum, such studies should include information on the distribution of GCS scores so that the validity of the comparisons can be judged.

For the future, further validation of the GCS is necessary. Even though further observer agreement studies are needed, they should be supplemented by studies that compare ratings with those by expert observers. For both types of studies, it seems essential that results be reported separately for those patients who are at intermediate levels of consciousness, and for those who are judged to be fully conscious or fully unconscious. Only by demonstrating high levels of accuracy for both patient conditions can the claims made about the reliability and validity of the GCS for the past fifteen years be fully supported.

We thank Ms Glenn Gardner and Ms Elizabeth Pittman (La Trobe University), and Ms Judy King and Ms Claire Warmuth (Alfred Hospital) for their help; Miss E. June Allen and Prof Judith Parker for promoting the Alfred Hospital-La Trobe University Collaborative Nursing Research Project; the patients and nurses for their contribution; and Prof Dennis Lowther, Dr Ian Mackay, Dr Merrill Rowley, and Dr Peter Tutton of the Monash University Medical Faculty for their advice.

REFERENCES

1. Teasdale G, Jennett B. Assessment of coma and impaired consciousness. *Lancet* 1974; ii: 81–84.
2. Teasdale G, Knill-Jones R, Van der Sande J. Observer variability in assessing impaired consciousness and coma. *J Neurol Neurosurg Psychiatry* 1978; **41**: 603–10.
3. Teasdale G. Assessing 'conscious level'. *Nursing Times* 1975; **72**: 914–17.
4. Jennett B, Teasdale G. Aspects of coma after severe coma head injury. *Lancet* 1977; i: 878–81.
5. Jones C. Monitoring recovery after head injury: translating research into practice. *J Neurosurg Nursing* 1979; **11**: 192–98.
6. Teasdale G, Gentleman D. The description of 'conscious level': a case for the Glasgow coma scale. *Scottish Med J* 1982; **27**: 7–9.
7. Allan D. Glasgow coma scale. *Nursing Mirror* 1984; **158**: 31–34.
8. Stanczak DE, White JG, Gouview WD, et al. Assessment of consciousness following severe neurological insult. *J Neurosurg* 1984; **60**: 955–60.
9. Starmark J, Stalhammar D, Holmgren E, Rosander B. A comparison of the Glasgow Coma Scale and the Reaction Level Scale (RLS85). *J Neurosurg* 1988; **69**: 699–708.
10. Fielding K, Rowley G. Reliablity of assessments by skilled observers using the Glasgow Coma Scale. *Aust J Adv Nursing* 1990; **7**: 13–17.
11. Cronbach LJ, Gleser GC, Nanda HK, Rajaratnam N. The dependability of behavioral measurements: theory of generalizabilty for scores and profiles. New York: Wiley, 1972.
12. Shavelson RJ, Webb NM, Rowley GL. Generalizability theory: new developments and novel applications. *Am Psychol* 1989; **44**: 922–32.
13. Lord FM, Novick ML. Statistical theories of mental test scores, Reading, Massachusetts: Addison-Wesley, 1968.
14. Ingersoll GL, Leyden DB. The Glasgow coma scale for patients with head injuries. *Crit Care Nurse* 1987; **7**: 26–32.
15. Starmark J, Holmgren E, Stalhammer D. Current reporting of responsiveness in acute cerebral disorders: a survey of the neurosurgical literature. *J Neurosurg* 1988; **69**: 692–98 (see p. 696).

VIEWPOINT

Prevention versus chemophobia: a defence of rodent carcinogenicity tests

PETER F. INFANTE

Anxiety about chemicals found to be carcinogenic in rodent studies has been labelled "chemophobia".[1] The spread of this phobia has been attributed to "phantom hazards" identified by current cancer testing methods.[2] If this argument is correct, public anxiety can be reduced; if not, arguments that the tests are meaningless may damage the long-term struggle to protect the health of the public. Cancer is no phantom; in the USA it affects more than one in four. In industrialised countries there have been increases in almost all forms of cancer over the past two decades in people over age 54 years (the ages at which most cancers occur).[3] Thus, it seems reasonable to suspect that environmental factors contribute to the rise in cancer incidence (and mortality).

The methodology that has been criticised[1,2,4] is the use of the "estimated" maximum tolerated dose (EMTD) as the high-dose level in cancer bioassays. Some writers have described the EMTD as a massive dose. It is not. The maximum tolerated dose is defined as "the highest dose of the test agent during the chronic study that can be predicted not to alter the animals' longevity from effects other than carcinogenicity . . .".[5] Moreover, the National Cancer Institute investigators who devised the cancer testing protocols by administering known human carcinogens to laboratory animals found that cancer developed only in the animals exposed to the EMTD.[6]

Some critics[2,4] conclude that the cancers that appear after the administration of a chemical at the EMTD are simply a reflection of increased cellular proliferation, not a result of a true carcinogenic response. According to these critics, a non-genotoxic substance given at the EMTD leads to such a high level of new cell proliferation that cell multiplication per se causes cancer. At lower doses, when there is no excess proliferation, there would be no cancer. This view, put simply, is that any substance given in a high enough dose becomes a carcinogen. For genotoxic carcinogens, the critics argue that the number of mutations leading to cancer response would vary at the same dose level depending on the amount of increased cellular proliferation induced by the test chemical at that dose level. The dose response below the level of increased cellular proliferation presumably would be

ADDRESS: Health Standards Program, Occupational Safety and Health Administration, Department of Labor, Room N3718, 200 Constitution Avenue, NW, Washington DC 20210, USA. (Dr P. J. Infante, Dr PH).