

A measure of neurobehavioral functioning after coma. Part I: Theory, reliability, and validity of the Disorders of Consciousness Scale

Theresa Louise-Bender Pape, DrPH, MA, CCC-SLP/L;^{1-4*} Allen W. Heinemann, PhD, ABPP;³⁻⁵
James P. Kelly, MA, MD;⁶ Anita Giobbie Hurder, MS;⁷ Sandra Lundgren, PhD, LP, ABPP⁸

¹The Department of Veterans Affairs (VA), Veterans Health Administration (VHA), Research Service, Edward Hines Jr. VA Hospital, Hines, IL; ²Marianjoy Rehabilitation Hospital, Wheaton, IL; ³Northwestern University, Feinberg School of Medicine, Department of Physical Medicine and Rehabilitation, ⁴Institute for Health Services Research and Policy Studies, Chicago, IL; ⁵Rehabilitation Institute of Chicago, Center for Rehabilitation Outcomes Research, Chicago, IL; ⁶University of Colorado School of Medicine, Department of Neurosurgery, Denver, CO; ⁷Cooperative Studies Program Coordinating Center, Edward Hines Jr. VA Hospital, Hines, IL; ⁸Minneapolis VA Medical Center, Mental and Behavioral Health Patient Service Line (M/C 116 B), Minneapolis, MN

Abstract—This longitudinal validation study describes the psychometric properties of the Disorders of Consciousness Scale (DOCS). This is Part I of a two-part series. Part II illustrates and describes the clinical and scientific implementation of the DOCS measure. The study was conducted at one intensive care unit, two acute rehabilitation hospitals, and one long-term acute chronic care hospital. Participants were unconscious after severe brain injury (BI). We conducted interrater reliability analyses using ratings from interdisciplinary pairs. Results indicated a higher-than-expected level of agreement and no significant difference between any pairs ($\chi^2 = 8_{5df}, p = 0.15$) (df = degrees of freedom). Examinations of ratings by discipline groups indicated that the DOCS is impacted minimally by discipline. Validity analyses demonstrate that 23 of 34 test stimuli remain stable over time with no floor or ceiling effect. DOCS measures obtained within 94 days of injury predicted recovery of consciousness up to 1 year after injury (c -indices of 0.70 and 0.86). Positive (0.71) and negative (0.68) predictive values indicate that the DOCS predicts recovery and lack of recovery. Twenty-three of the DOCS test stimuli produce a reliable, valid, and stable measure of neurobehavioral recovery after severe BI that predicts recovery and lack of recovery of consciousness 1 year after injury.

Key words: brain injury, coma, consciousness, measure, outcome, psychometrics, recovery.

Abbreviations: BI = brain injury, CHI = closed-head injury, CI = confidence interval, CRS = Coma Recovery Scale, df = degrees of freedom, DOCS = Disorders of Consciousness Scale, GCS = Glasgow Coma Scale, ICU = intensive care unit, LOSIPR = length of stay for inpatient rehabilitation, MCS = minimally conscious state, PCA = principal component analyses, ROC = receiver operating characteristic, SD = standard deviation, SMART = Sensory Modality Assessment and Rehabilitation Technique, VA = Department of Veterans Affairs, VS = vegetative state, WNSSP = Western Neuro Sensory Stimulation Profile.

This material was based on work supported by the Department of Veterans Affairs, Veterans Health Administration, Rehabilitation Research and Development Service, through a career development grant to Dr. Pape (B2632-V) and through the Health Services Research and Development Service, Midwest Center for Health Services and Policy Research (locally initiated project 42.063). Funding was also provided by the U.S. Department of Education, National Institute on Disability and Rehabilitation Research, through Advanced Rehabilitation Research Training Program grant CFDA 84.133P and a Merit Switzer Award to Dr. Pape (CFDA 84.133F).

*Address all correspondence to Dr. Theresa Louise-Bender Pape, Department of Veterans Affairs (VA), Veterans Health Administration, Research Service, Edward Hines Jr. VA Hospital, PO Box 5000 (M/C 151H), Hines, IL 60141; fax: 708-202-7487; email: Theresa.Pape@med.va.gov
DOI: 10.1682/JRRD.2004.03.0032

INTRODUCTION

Severe brain injury (BI) results in sudden altered consciousness of varying duration [1]. This state of altered consciousness has been described by two subsyndromes, coma and vegetative state (VS) [2–3]. A third subsyndrome, the minimally conscious state (MCS), was defined in 1996 as a transitional state indicating either improvement in consciousness or deterioration in level of consciousness (**Appendix, Table 1**, available in online version only) [4–7]. Though these terms are used to describe a continuum of altered consciousness, no tool adequately describes changes in functioning. Clinicians need an assessment tool that—

1. Can be completed at bedside.
2. Is sensitive to subtle changes in neurobehavioral functioning.
3. Produces a reliable and valid measure of neurobehavioral functioning over time while the patient is unconscious.
4. Improves outcomes prediction.

Scientists need a reliable and valid measure of neurobehavioral functioning in unconscious persons to identify the factors that influence recovery and to examine the effectiveness of medical and rehabilitation interventions. The Disorders of Consciousness Scale (DOCS) was designed to address these clinical and scientific needs. Part I of these papers describes the development, purpose, and psychometric properties of DOCS. Part II (this issue, page 19) reports the sensitivity of DOCS and the implementation of DOCS in clinical and scientific practice.

Development and Purpose

The first version of the DOCS, developed between 1991 and 1992, was titled “Standardized Assessment of Consciousness.” The name was changed in 1995 to DOCS, and it was pilot-tested from 1992 through 1999 [8]. Development of the DOCS has been an iterative process, with pilot findings serving as the basis for revisions, including changes to the rating scale and test stimuli [9]. The reliability and validity of this refined version are summarized here.

The DOCS is a neurobehavioral bedside evaluation. What distinguishes the DOCS from other tools is that it was designed to measure neurobehavioral integrity from the perspective that—

1. The state of altered consciousness is a continuum.

2. A finite set of prescribed or expected responses does *not* serve as exhaustive indices of neurobehavioral functioning.
3. Our ability to monitor neurobehavioral recovery or change after severe BI is related to our ability to measure the amount or level of neurobehavioral functioning within the continuum of altered consciousness.
4. A sensitive, reliable, and valid measure of neurobehavioral functioning must maintain its meaning over time.

The DOCS test stimuli, administration procedures, and scoring procedures were designed to allow the clinician to examine unconsciousness as a continuum of fluctuating levels of neurobehavioral integrity while detecting and distinguishing between true change and random fluctuation.

The DOCS is different from other assessment tools, such as the Coma Recovery Scale (CRS) [10] and the Western Neuro Sensory Stimulation Profile (WNSSP) [11], in that the rating scale of the DOCS provides a description of neurobehavioral recovery. The rating scale describes levels of neurobehavioral integrity, and a level is assigned to responses to test stimuli, whereas CRS specifies the behavioral responses that are expected when a patient is given a test stimulus. If the patient does not demonstrate the behavior specified in the CRS, then the patient is assigned a lower score indicating less or no neurobehavioral functioning. The dichotomous data obtained from CRS reflect either the presence or absence of a specific behavior rather than the level of neurobehavioral functioning manifested.

The WNSSP was one of the first instruments designed to detect subtle changes in neurobehavioral functioning in low-level neurological states. The WNSSP test stimuli were a starting point for development of the DOCS test stimuli but were expanded and refined because the WNSSP test stimuli do not target lower-functioning patients [12]. While WNSSP and DOCS test stimuli are similar, test stimuli administration and scoring procedures are different. The WNSSP allows for cues and specifies that lower scores should be assigned if a patient responds to a test stimulus when provided with a cue and when response is delayed. Cueing techniques do facilitate behavioral responses and function, but the use of cues makes determining the amount of neurobehavioral functioning without priming impossible. The timeliness of responses to test stimuli is handled differently with DOCS, where procedures allow patients 10 to 30 seconds (depending on the stimuli) to respond. This procedure was implemented to discriminate responses to

test stimuli from random responses. The Sensory Modality Assessment and Rehabilitation Technique (SMART) is a relatively new instrument that distinguishes five levels of neurobehavioral functioning by consistency of behavioral responses [13–14]. A comparison of the measurement properties of the DOCS with those of CRS, WNSSP, and SMART reported in published literature is available in the **Appendix, Table 2** (available in online version only).

Development of Rating Scale, Test Stimuli, and Administration and Scoring Procedures

The DOCS comprises a baseline observation protocol, a three-point rating scale, and test stimuli. The baseline protocol provided in **Appendix, Table 3** (available in online version only), is completed prior to the examiner administering test stimuli, and the rating scale is used by the examiner to assign a level of neurobehavioral integrity to responses elicited with test stimuli.

DOCS Rating Scale: What Do the Raw Scores Mean?

Examiners use the DOCS rating scale to assign a score of 0, 1, or 2 to behavior(s) elicited with a test stimulus. A higher score indicates a higher level of neurobehavioral integrity. Multiple responses can indicate neurobehavioral integrity, but only the best response is used for computing the DOCS measure of neurobehavioral functioning. The original rating scale distinguished five levels (0-1-2-3-4) of neurobehavioral integrity [15] but was collapsed in 1999 to a three-category scale because not all rating scale points were used: 0 = No Response, 1 = General Response, and 2 = Localized Response [16].* The rating scale defines transitions from low to middle to high neurobehavioral functioning within the continuum of altered consciousness.

The DOCS comprises two scoring forms. Form B was developed in 1992 and includes the baseline observation protocol, stimuli administration procedures, and response interpretation guidelines. In 1999, Form B was expanded to also include examples, within each subscale, of behaviors that constitute general and localized responses. Form A, the

short version, was also developed in 1999 and includes the baseline observation protocol and scoring grids. Therapists choose either Form A or B, but a novice therapist is encouraged to use Form B.

DOCS Baseline, Test Stimuli and Administration and Scoring Procedures

The test stimuli are organized in eight subscales including Social Knowledge, Taste and Swallowing, Olfactory, Proprioceptive and Vestibular, Auditory, Visual, Tactile, and Testing-Readiness. The test items, in each subscale, are ordered from easy to difficult, and this ordering was based on pilot data [16].

Three ideas guided the development of the administration procedures and the selection of test stimuli. First, a method must exist for discriminating between true and random responses. The baseline observation protocol was developed as one means of addressing this issue. Test stimuli can only be administered after completion of the baseline protocol. The baseline observation protocol is a systematic checklist that is completed by the examiner observing the patient at rest. It takes 2 to 5 minutes to complete.

The second idea was that the administration procedures should reflect allied health clinical judgment. The procedures specify, for example, that easier items can be skipped if the examiner determines that a patient's ability exceeds the challenge presented by a given item. "Juice," for example, is the easiest item in the Taste and Swallowing subscale. "Massage," "SpoonW," "SpoonC," and "Tap" follow. If the examiner has previously observed the patient to lower his or her lips when presented with a cold spoon, then the first item at the patient's ability level would be "SpoonW." If the patient receives a score of "2" on the first item, then the examiner can skip the easier items within that subscale. If the patient scores a 1, then the examiner administers all easier items within that subscale. Subscale and test stimuli administration procedures are summarized in the **Appendix, Table 4** (available in online version only).

The third idea was that potential confounders to distinguishing between true and random responses should be controlled before the examiner administers the first test item and throughout the entire test. The procedures for controlling confounders include environmental (e.g., avoidance of extreme insults to the sensory system such as bright lights and unpredictable noises), positioning [17], and testing-readiness controls. Testing does not start until

*General Response = a response not related to the spinal tract and that differs from baseline behaviors that are either a reflex or a response not contextually related to the test stimuli. Localized Response = a contextually related response that differs from baseline behaviors and reflects an ability to regulate incoming sensory information, which is constantly changing, and to control responses to the sensory input.

environmental controls are in place. General positioning guidelines are to be followed throughout the evaluation, along with additional specifications for some test stimuli. The general guidelines describe positioning for lying in bed, sitting on the side of a bed or on the side of a mat, and sitting in a chair. These guidelines also specify that testing should be paused when a patient slips out of position. Pausing and repositioning allow the examiners to associate behavioral responses to test stimuli rather than positional pain. Some items, such as in the Taste and Swallowing subscale, specify further that the patient should be upright between 45° and 90° with his or her head and neck at midline and supported. Testing-readiness is defined as a general state of readiness to respond, and it is observed and measured behaviorally. The testing-readiness controls include procedures for describing a behavior used to indicate a state of readiness. The testing-readiness procedures are completed by the examiner immediately after baseline observations and before administering the first test stimulus. Testing-readiness is reestablished if the patient demonstrates a reduced state of readiness. A separate subscale, called testing-readiness, is used to track the amount and type of stimulation provided to reestablish this state of readiness.

METHODS

Ninety-five persons aged 18 years and older with severe BI (Glasgow Coma Scale) [$GCS \leq 8$] were recruited from one intensive care unit (ICU), two inpatient (IP) rehabilitation hospitals, and one long-term acute chronic hospital in the Midwestern United States. Persons with closed- and open-head injuries and anoxia were included. Patients were excluded if the referring hospital did not calculate the GCS score before administration of neuroparalytic agents or at the half-life of these agents [18]. Also excluded were patients with histories of neurological and/or psychological disorders. Informed consent was obtained from legal representatives. If research participants recovered consciousness during their 1-year participation, healthcare providers evaluated the participant's healthcare decision-making capacity. Participants demonstrating capacity were reconsented [19]. The human subject institutional review boards at the participating hospitals approved the study.

Instrumentation and Data Collection Procedures

Each participant in this study was evaluated weekly with the DOCS—up to 6 weeks. DOCS evaluations were discontinued when a participant met criteria for having recovered consciousness. The time point during the recovery continuum when the first DOCS test was completed was based on clinical considerations. Participants were assessed as early as 8 days after injury and as late as 424 days after injury. The best response to each DOCS test item was scored. Complete DOCS evaluations required 45 to 60 minutes to complete unless starting and skipping rules were followed, which reduced administration time to 30 minutes. For this study, examiners administered as many items as possible within 1 hour. Examiners completed each DOCS assessment individually or in interdisciplinary pairs after completing 2 hours of training. Examiners completed 383 evaluations, with each evaluation comprising 34 test stimuli, resulting in 13,022 ratings of which 9,892 (76%) were conducted in interdisciplinary pairs. During IP rehabilitation, each participant was screened three times a week for indications of consciousness. The screenings occurred during routine therapy sessions during IP rehabilitation and routine medical procedures in the ICU. After IP rehabilitation, clinical research personnel conducted monthly screenings up to 1 year to identify when or if the participant recovered consciousness. Recovery of consciousness was defined as demonstrating one of the following: (1) functional interactive communication, (2) functional use of an object, or (3) a behavioral manifestation of sense of self in an environment that can be documented. When a participant was screened for indications of consciousness and judged to be more responsive than indicated by his or her behavior during the screening, then the examiner informally chatted with the participant. The examiner, for example, may have judged the participant to be bored with the activity. The examiner may have, therefore, told the participant a silly joke. If the participant laughed in a contextually appropriate manner, then he or she was rated conscious because the examiner provided a written description of the behavior. If this situation arose after IP discharge, then the consciousness screenings were conducted face-to-face during routine outpatient clinic visits.

Transformation of Behavioral Data: Raw Scores to Logits to DOCunits

Therapists administered DOCS items and scored behavioral responses, which differed from baseline

behaviors, as 0 (No Response), 1 (General Response), or 2 (Localized Response). These scores were then converted to an equal-interval measure with the use of the rating scale model [20] and facets model [21].* This conjoint (additive) probability model estimates person measures and item difficulties with the use of the maximum likelihood estimation [22] for each element specified in the models (i.e., person ability, test item difficulty, and rater severity) [23–25].

Given that the range of the DOCS instrument, based on the first DOCS evaluation (DOCS-1), is approximately 8 logits (–4.0 to 4.0, **Figure 1**), the logit scale can be transformed to a scale that is more easily understood [DOCunit = 50 + (logit × 12)]. This convenient transformation is referred to, from this point forward, as the “DOCunit” and gives the DOCS a range of 0 to 100 (**Figure 2**). The slight differences apparent between **Figures 1** and **2** are due to the choice of interval endpoints, not the scaling change. After this transformation, the standard error of the DOCS measure for a participant with all 34 items administered is approximately 4 DOCunits. With this precision, decimal places are uninformative at the individual level and are not reported. The conversion of raw scores to logit and then to the DOCunit allows DOCS measures to be easily understood and used in parametric statistics.

Data Analyses: Reliability and Validity

The rating scale model provides estimates of separation reliability, interrater agreement/reliability, rater severity, and construct validity. We used the rating scale model to analyze the stability of the rating scale over time, the fit of each test item to the underlying construct of neurobehavioral functioning, and the fit of each participant to the response sets of the entire sample. It was also used to examine the stability of item calibrations over time. We used the facets model to examine interrater reliability and rater severity because it is in the form of a logistic regression model, but each person, item, and rater is individually parameterized. DOCS measures derived from the previously described transformations are used

*Rating scale model: $\log[P_{nik}/P_{ni(k-1)}] = B_n - D_i - F_k$ and facets model: $\log[P_{nij}/P_{nij(k-1)}] = B_n - D_i - C_j - F_k$, where B_n is the ability of each participant, D_i is the difficulty of each test stimulus, C_j is the severity of the rater (therapist), and F_k is the calibration measure of rating category k relative to $k - 1$; difficulty attributed to transitioning from one step in the rating scale to the next (0 transitioning to 1 transitioning to 2).

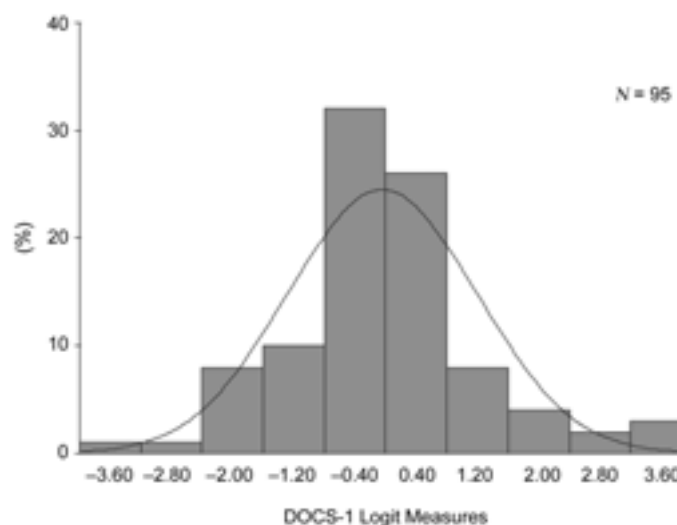


Figure 1. Distribution of initial DOCS measures (DOCS-1) for total sample in logit scale.

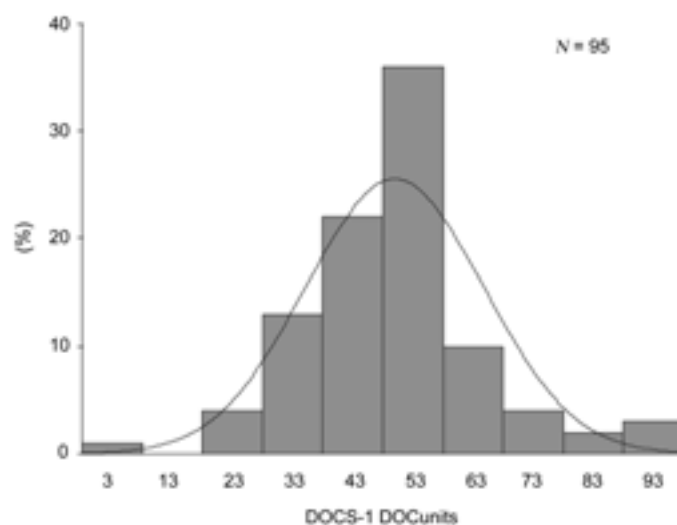


Figure 2. Distribution of initial DOCS measures (DOCS-1) for total sample in DOCunits.

as point estimates in the bivariate and multivariate analyses of predictive validity.

RESULTS

The sample of 95 participants was largely composed of young (mean age at injury = 36 years) white (73%)

males (85%) with closed-head injuries (CHI) (72%) (Table 1), who at the time of injury were either married (45%) or single (45%), had received some college education (34%), were employed full-time (53%), lived in a

household with an annual income \geq \$50,000 (60%), and were insured through a preferred provider organization (40%). Of these participants, 22 percent are eligible for veteran healthcare benefits.

Table 1.

Demographics at time of injury: Total sample, CHI, and other BI.

Variable	All BI (N = 95)	CHI (n = 68)	Other BI (n = 27)	Sample Sizes
Age (Mean \pm SD)	36 \pm 15	35 \pm 16	40 \pm 14	—
Race (N = 95)				
White	69 (73%)	56 (82%)	13 (48%)	69
Black	16 (17%)	7 (10%)	9 (33%)	16
Other	10 (10%)	5 (8%)	5 (19%)	10
Gender (N = 95)				
Male	81 (85%)	59 (87%)	22 (81%)	81
Female	14 (15%)	9 (13%)	5 (19%)	14
Marital Status* (N = 94)				
Married	42 (45%)	27 (40%)	15 (55%)	42
Single	42 (45%)	34 (51%)	8 (30%)	42
Divorced or Separated	9 (10%)	5 (8%)	4 (15%)	9
Widowed	1 (<1%)	1 (1%)	0 (0%)	1
Education* (N = 86)				
\leq Grade 11	9 (10%)	6 (9%)	3 (14%)	9
High School or GED	21 (24%)	18 (28%)	3 (14%)	21
Some College (w/o degree)	29 (34%)	21 (33%)	8 (36%)	29
Community College or Trade School Degree	11 (13%)	6 (9%)	5 (22%)	11
Bachelors and/or Graduate Degree	16 (19%)	13 (21%)	3 (14%)	16
Employment* (N = 87)				
Unemployed	20 (23%)	14 (22%)	6 (28%)	20
Full-Time	46 (53%)	36 (55%)	10 (45%)	46
Part-Time	13 (15%)	9 (14%)	4 (18%)	13
Full-Time Student	8 (9%)	6 (9%)	2 (9%)	8
Insurance* (N = 81)				
Uninsured	6 (7%)	6 (10%)	0 (0%)	6
HMO	13 (16%)	9 (16%)	4 (17%)	13
PPO	32 (40%)	21 (36%)	11 (48%)	32
Private Pay	13 (16%)	10 (17%)	3 (13%)	13
Other	17 (21%)	12 (21%)	5 (22%)	17
Household Income* (N = 75)				
\leq \$14,999	11 (15%)	9 (16%)	2 (11%)	11
\$15,000 to \$49,999	19 (25%)	14 (25%)	5 (26%)	19
\geq \$50,000	45 (60%)	33 (59%)	12 (63%)	45

*Sums do not reach total sample sizes of 95, 68, and 27 because of missing data when cross tabulations are completed.

SD = standard deviation, GED = general equivalency diploma, PPO = preferred provider organization, HMO = health maintenance organization, w/o = without, All BI = all brain injuries regardless of etiology, CHI = closed-head injury, Other BI = other types of brain injury (anoxic, aneurysm, open-head injury, arteriovenous malformation, and one hemorrhage)

At time of injury, CHI participants (68/95) and other BI participants (27/95) were similar in age, gender, marital status, educational level achieved, employment status, and household income. The two groups significantly differ in proportion of race ($\chi^2 = 11.375_{1df}$, $p = 0.001$) such that nonwhite participants represented 19 percent of the CHI sample and 52 percent of the other BI sample.

The average duration of acute rehabilitation for the total sample was 51.5 days ($n = 88^*$) days. The average length of stay in acute rehabilitation is not significantly longer for participants with CHI (51 days) compared with participants with other BIs (54 days). Each participant received an average of 113.0 hours of rehabilitation services and an average of 2.5 hours a day of acute IP rehabilitation over a 7-day work week. Participants with CHI did not significantly differ from other BI participants according to rehabilitation intensity.

Examination of DOCS Rating Scale

For all participants ($N = 95$), the DOCS rating scale reflects progressively improving levels of functioning as demonstrated by the monotonic ordering of the average DOCunit measures for each category of the rating scale (0 = -8.0, 1 = 0.10, 2 = 8.5). This indicates that lower-rating categories were more probable for persons with lower levels of neurobehavioral functioning and the higher-rating category was more probable for persons with higher levels of neurobehavioral functioning. Transition points between categories of the rating scale, step threshold measures, are also monotonically ordered (-15.71, 15.71), indicating that each of the three rating categories is most likely to be used according to improving status. Scale stability is also evidenced by the observation that the majority of the items (76%, 26/34) and the corresponding average measures for each of the 34 items, according to each category of the rating scale, maintain monotonic ordering (**Appendix, Table 5**, available in online version only) [26].

Examination of Interrater Reliability: Agreement and Severity

Allied health professionals who conducted DOCS evaluations included 12 speech-language pathologists, 12

physical therapists, 14 occupational therapists, 2 registered nurses, 2 neuropsychology doctoral candidates, and 2 respiratory therapists. We examined the manner in which these allied health professionals rated behavioral responses to determine if differences individually and by discipline groups affect the DOCS measure. We examined reliability of raters by computing the percentage of exact agreement and by comparing the observed with the predicted agreement. For example, if a speech pathologist had given 5,123 ratings, then we would have examined the ratings of all the other raters to determine if any had been given under identical circumstances (i.e., same person, item, and task). If a match was found, then this would have been an “exact agreement opportunity.” This procedure is repeated for all the other ratings and raters. Over the entire data set, we found 33,003 exact agreement opportunities. The percentage of actual exact agreements under identical conditions (54.4%) is slightly greater than the percent agreement predicted by the facets model (43.8%).[†] This finding indicates that the raters are acting as independent experts and are unlikely to be rating by consensus. The ratings between all pairs are not significantly different ($\chi^2 = 8_{5df}$, $p = 0.15$), suggesting that there is a higher-than-predicted level of agreement between all the pairs of raters.

In addition, we examined individual raters according to rating pairs and according to allied health disciplinary groups. Findings indicate that the DOCS measure is impacted according to discipline group by only 0.18 raw score points (**Table 2**) as evidenced by the range of adjusted averages across discipline groups (0.18 = 1.22 – 1.04). Neuropsychology raters were the most lenient but differed from speech pathology raters by only 0.15 raw score points on any given behavioral response. Collectively, these findings indicate that the raters are rating in the same manner and that the impact of rater leniency or severity on the DOCS measure is minimal.

[†] $\kappa = (\% \text{ observed agreement} - \% \text{ expected agreement}) / (100\% - \text{expected agreement}) = (0.544 - 0.438) / (100 - 0.438) = 0.001$; conventionally, the “expected agreement %” is the level of chance agreement based on the marginal frequencies of the contingency tables. Values of κ above 0 are desired. But, under Rasch model conditions, the “expected agreement %” is the model prediction, and so the expected value of κ is 0.0.

*Rehabilitation utilization data are derived from billing records. VA hospitals do not have billing procedures. Data are not available for veterans.

Table 2.

Rater severity for total sample by allied health discipline.

Rater Group	Observed Raw Score	Observed Count	Observed Average	Outfit Mean Square
SLP	4,976	5,123	1.0	1.0
PT	1,707	1,779	1.0	1.0
OT	1,676	1,729	1.0	1.1
RN	604	665	0.9	1.0
NP	86	75	1.1	0.9
RT	548	521	1.1	1.0

SLP = 12 speech-language pathologists, PT = 12 physical therapists, OT = 14 occupational therapists, RN = 2 registered nurses, NP = 2 neuropsychology doctoral candidates, RT = 2 respiratory therapists

Observed raw score = observed raw score, sum of raw scores for total sample.

Observed count = number of active responses.

Observed average = (observed score/observed count).

Outfit mean square = outlier sensitive mean square fit statistic, with expectation 1, and range of 0 to infinity. It is standard chi-square ÷ its degrees of freedom.

Examination of Construct Validity

Evidence of construct validity is provided by how well the DOCS test measures what it purports to measure (neurobehavioral functioning). If the responses describe neurobehavioral functioning meaningfully, then MCS participants should manifest more localized responses while VS participants should demonstrate more generalized responses to the difficult items. Demonstrating localized responses to each incrementally more difficult task should translate to more intact central nervous system processing. Construct validity is evaluated with principal component analyses (PCA) of residuals and with the examination of fit indices and item calibrations for each time point.

We conducted PCA of item residuals to determine whether a secondary dimension is in the test items or whether the unexplained residual variance can be attributed to random fluctuations in the observations. PCA detected correlations among the item residuals. Results indicate that the DOCS measure (eigenvalue = 53.5) explained the majority (53.5/87.5, 61%) of the total variance in the observations; the first factor of the residuals accounted for only 4 percent of the residual variance (eigenvalues = 3.5/34.0). Comparing the strength or power of the 34 DOCS items to the power of the first factor allows for a determination of whether 4 percent is or is not a meaningful secondary dimension. Eigenvalue for the 34 DOCS items is 15 times stronger than the eigenvalue for the first factor, suggesting that the structure to the unexplained variance in the item residuals is negligible. An additional examination of the factor contrasts confirms that no meaningful substructure exists. Together, this evidence indicates that the first factor in

the residuals is dominated by noise, and there is no practical impact on the measurement of neurobehavioral functioning with the DOCS test items.

We conducted PCA of the residuals for each participant's DOCS measure to determine whether the sample represents one dimension of severe BI or whether analyses should be stratified by subsamples. Results indicate that the estimated level of neurobehavioral functioning for each participant (DOCS measure) explains the majority of the total variance (61%; eigenvalues = 401.5/656.5, respectively) and that the first factor explained 4 percent of the total unexplained variance (eigenvalues = 23.6/255.0). Comparing the power of the DOCS measure to the power of the first factor indicates that the DOCS measure is 17 times stronger. This comparison suggests that the structure in the person residuals is negligible, but additional examination of the factor contrasts suggests that while the structure is negligible, it may be clinically meaningful if examined by etiological subgroups. That is, the majority of the participants with other BI (78%, 21/27) fell on one end of the severe BI dimension and persons with CHI fell on the other end of the dimension. The six participants with other BI who had factor loads similar to CHI participants incurred injuries due to anoxia, gunshot wounds, or falls resulting in skull fractures. The items most sensitive to this contrast are "Juice" and "Focus." Together, the evidence suggests that the sample is not substantively composed of different persons, but the sample *may* be composed of different types of injuries. Analyses to construct the DOCS measure described in the following paragraph were, therefore, conducted for the total sample and by subsamples (i.e., CHI and other BI).

We further examined construct validity by analyzing fit indices for each item by time and by examining item calibrations according to time. Time 1 means assessment number 1, and Time 2 means assessment number 2, etc. We obtained item fit indices and item calibrations for Times 1 through 6 by holding the Time 1 person mean constant during estimations of fit statistics and calibrations (i.e., racked data). These racked analyses allowed for examining fit statistics for each item at each time point and for identifying the item calibrations that changed from Time 1.

An examination of the fit statistics obtained, with the use of the procedures just described, indicates that items do not overfit (mean square ≤ 0.70) and are not overly predictable until the final time point (Time 6). This finding is as expected because as participants begin to recover, they begin to respond to test stimuli in a more predictable and consistent manner. The fit range applied in this examination (i.e., acceptable mean square range of 0.7 to 1.3) is more conservative than the range recommended for observational data (i.e., mean square range of 0.5 to 1.7) and indicates that 25 of the 34 test items fit the underlying construct for both samples [27–28]. More details about item fit statistics by samples and time can be found in the **Appendix, Table 6** (available in online version only).

We examined item calibrations obtained (using the previously described procedures) for each time point by plotting item calibrations from Time 1 versus 2, 1 versus 3, 1 versus 4, 1 versus 5, and 1 versus 6 for each sample. We identified 11 items as unstable (e.g., “Tap” and “Stroke”). For further details, see **Appendix, Table 7** (available in online version only). The calibrations for these 11 items changed between Time 1 and Time 6 from 10 to 25 DOCunits and fell outside the bounds of the 95 percent confidence interval (CI) when examined with the use of subsamples (CHI versus other BI). These 11 items were, therefore, eliminated. After eliminating these 11 items, we reexamined item calibration stability using the same procedures. We examined the remaining 23 items by plotting item calibrations in the same manner for each sample (**Figure 3**). The remaining 23 items fall within the upper and lower bounds of the 95 percent CI for the CHI and other BI samples. Since the person mean was held constant in these analyses, the fact that the trend line is below the identity line in **Figure 3(a)** and **(b)** is a reflection of the improvement for the entire sample between the first assessment and the third assessment. The outfit mean square statistics and item calibrations for each of these 23 items by time and sample can be found

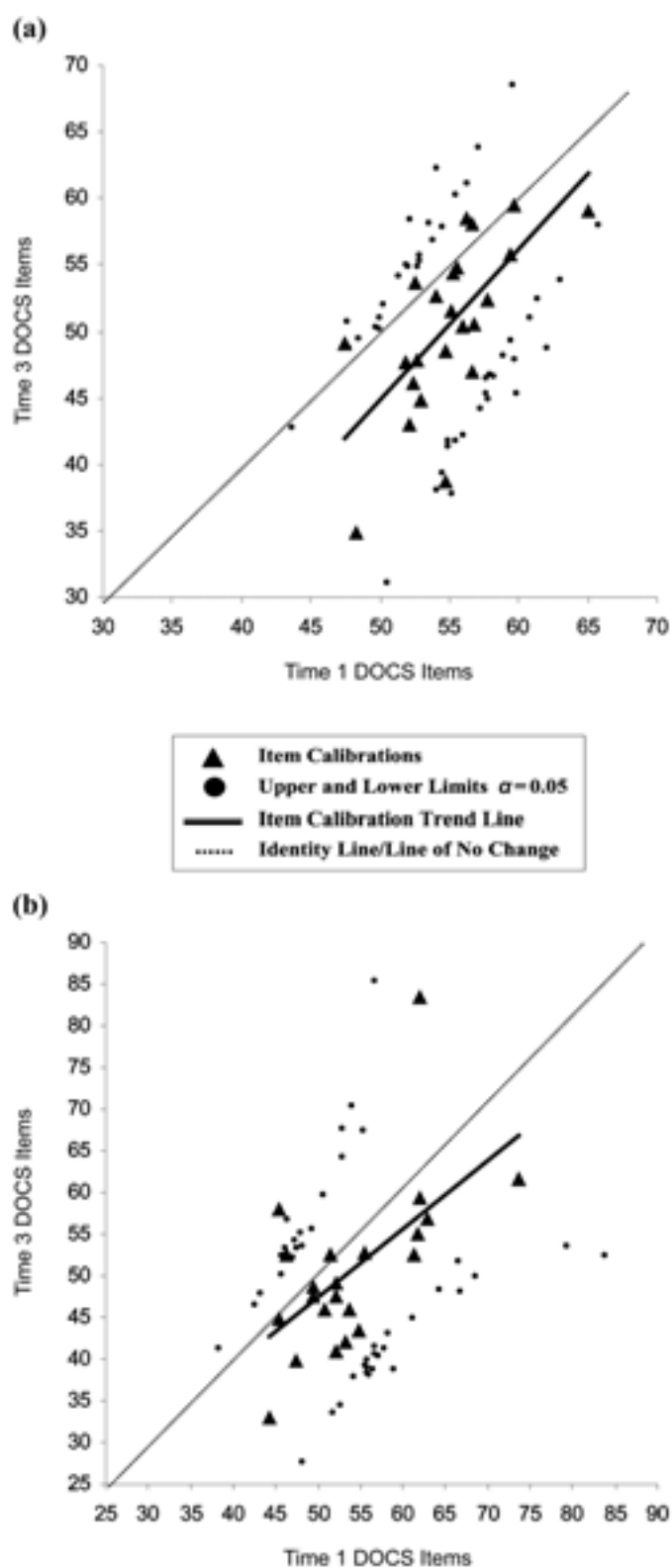


Figure 3. Item calibration stability by time and samples: (a) closed-head injury and (b) other brain injury samples.

in the **Appendix, Tables 8 and 9**, respectively (available in online version only).

A reexamination of the outfit statistics for the remaining 23 items indicates that all items fall within the acceptable range of 0.70 to 1.30 (**Table 3**) and provide independent information about neurobehavioral functioning. While the item calibrations for these 23 items remain stable over time for both samples, it is important to note that the item calibrations are different according to the samples (**Figure 4**). The plot in **Figure 4** illustrates that all but two of the items (“Focus” and “Air”) fall along a diagonal. The slightly different calibrations explain the different item ordering for each subsample as shown in **Table 3** and suggest that the items measure different aspects of neurobehavioral functioning for each sample. This finding confirms the results from the PCA of person

residuals and indicates that all future calibrations to compute the DOCS measure should be stratified by etiology.

Targeting the Test to the Samples

The targeting of the items to the sample is evidenced through the comparison of the average person measure with the average item calibration measures. The average person measure for CHI is 50.31 (standard deviation [SD] = 11.33) DOCunits and the item mean is 50.00. The average person measure for the other BI is 46.45 DOCunits (SD = 8.05) logits, and the item mean is 50.00. The person means, for both samples, are within 5 DOCunits of the item means. A comparison of the ranges and the averages indicates that the DOCS test is targeted to persons who are recovering from coma after CHI and other BI. There are test items that challenge persons who are comatose, vegetative and minimally conscious.

Table 3.

Average item calibrations in DOCunits and outfit mean squares by difficulty and samples for remaining 23 items.

CHI Sample					Other BI Sample						
Item No.	Item Name	Description	Item Calibration	Outfit Mean Square	Item No.	Item Name	Description	Item Calibration	Outfit Mean Square		
T3	HAIR	Hard	61.3	0.97	T3	HAIR	Hard	60.9	0.07		
C1	GREET	↑ ↓	56.7	1.28	V5	TRACKING	↑ ↓	59.3	0.99		
T7	SWAB		56.1	1.20	T7	SWAB		58.1	0.79		
T6	SCRUB		55.3	1.08	V7	TRAKFACE		57.2	0.90		
T1	AIR		53.2	1.02	V4	FOCUS		56.7	0.85		
V5	TRACKING		53.1	0.74	V8	FOCUSFAC		51.9	0.14		
V7	TRAKFACE		52.5	0.73	C1	GREET		51.7	0.85		
A5	BELL		52.1	0.81	A6	COMMAND		51.7	0.11		
T5	HAND		51.7	0.80	A5	BELL		51.2	0.72		
A3	NAME		50.7	0.60	T6	SCRUB		50.6	0.76		
A6	COMMAND		50.2	0.77	T5	HAND		50.5	0.20		
A1	WHISTLE		49.5	1.07	PV1	JOINT		49.8	0.10		
T2	FEATHER		48.6	0.87	A3	NAME		48.4	0.73		
S2	MASSAGE		47.6	1.71	T8	CUBE		47.9	0.15		
A2	CLAP		47.4	0.99	V3	BLINK		47.8	0.06		
V4	FOCUS		47.3	0.84	O1	ODOR		47.6	0.26		
T8	CUBE		47.2	1.01	T2	FEATHER		46.8	0.67		
V8	FOCUSFAC		47.0	0.79	S2	MASSAGE		46.6	0.13		
O1	ODOR		46.8	1.02	T4	TOE		46.3	0.35		
T4	TOE		46.3	1.24	A1	WHISTLE		44.9	0.15		
PV1	JOINT		45.3	0.94	T1	AIR		43.6	0.97		
V3	BLINK		Easy	44.2	1.51	A2		CLAP	Easy	41.9	0.00
S1	JUICE		40.0	1.40	S1	JUICE		38.6	0.10		
—	—		MEANS	50.0	1.0	—		—	MEANS	50.0	0.40

CHI = closed-head injury, Other BI = other types of brain injury

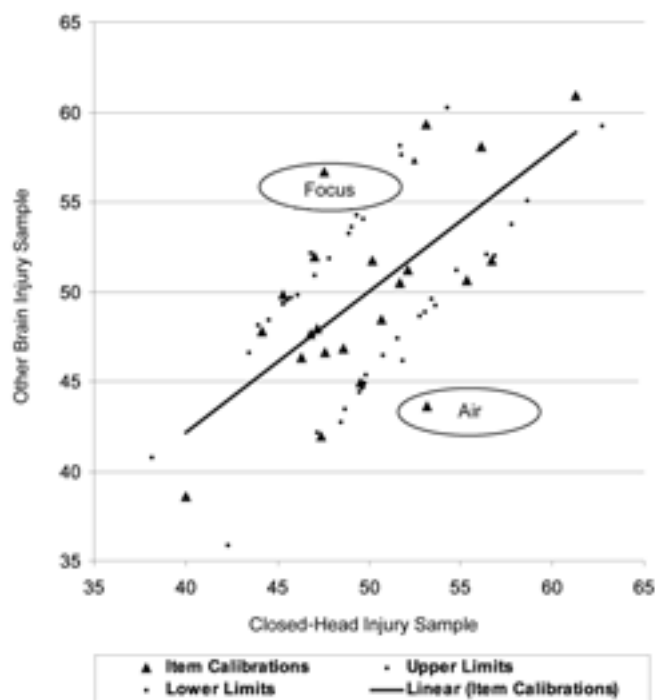


Figure 4. Item calibration for 23 remaining DOCS items by closed-head injury and other brain injury samples.

The person separation reliability of the DOCS measure, built on the refined set of 23 items, illustrates the robustness and sensitivity of the final DOCS measure. Person separation reliability indices of 2.38 for CHI (Cronbach's alpha = 0.85) and 1.81 for OBI (Cronbach's alpha = 0.77) indicate that the items reliably differentiate three levels of neurobehavioral functioning [29–31].

Predictive Validity

Predictive validity is examined with bivariate analyses, mixed random effects regression analyses, a comparison of four logistic regression models, and a comparison of the actual versus predicted outcomes. We examined 13 predictor variables (Table 4). DOCS measures derived from the refined set of 23 DOCS test items are the primary predictor variables of interest and were used as point-estimates in validity analyses. The dichotomous outcome is whether or not a participant recovered consciousness within 365 days of injury.

DOCS-Slope and DOCS-Intercept: Time Along the Recovery Continuum

DOCS-Slope and DOCS-Intercept are 2 of the 13 predictor variables, and we estimated them using a mixed

random effects regression model. This model was used because the number of DOCS measures for each participant is unequal and because the participants are all measured at different time points during the recovery continuum (8 to 424 days after injury; while each participant was followed for 365 days, the final interviews for some participants required additional days to schedule the interview). Mixed random effects regression modeling uses empirical bayes estimation to estimate each participant's intercept and recovery slope [32]. We computed the DOCS-Intercept using each participant's DOCS-1 measure, and after controlling for type of injury (CHI or other) and time after injury (i.e., DOCS-1 Days), we found that DOCS-1 measure reflects initial neurobehavioral severity. The DOCS-Slope reflects each participant's recovery rate and comprises at least 1 and up to 6 DOCS measures.

Given the wide range of time after injury when DOCS assessments are completed, time is treated as a random variable, but it is made more uniform with the assignment of each DOCS measure to one of nine time categories. Time 0 (intercept) includes all DOCS measures derived between 7 and 21 days after injury. Time 1 includes all DOCS measures derived between 22 and 43 days. Time 2 reflects all measures derived between 44 and 65 days after injury. Times 3, 4, 5, and 6 reflect all measures derived between 66–87, 88–109, 110–131, and 132–153, respectively. Time 7, represents all DOCS measures derived between 154 and 180 days after injury. The final time category, Time 8, includes all measures derived after 180 days. This categorization was done for mixed random effects regression analyses only, and time remains uncategorized for all other validity analyses.

Mixed Random Effects: Initial Severity and Recovery Rates by Individuals and Groups

For the mixed random effects regression model, the individual participants and time (time categories 0 through 8) served as random effects. The fixed effects were the etiological group ($N = 95$; CHI = 68, other BI = 27) and time by group interaction (Time \times Group). Results indicate that the CHI and other BI groups did not significantly differ according to initial severity (mean DOCS = 43.04 DOCunits). Both groups exhibited an overall improvement of 51.08 DOCunits every 3 weeks (21 days). This finding means a statistically significant change takes about 6 months, but a clinically significant change of 50 DOCunits takes about 4 months. The rate of improvement between the two groups was not significantly different, but a significant variation was found in

Table 4.

Predictor variables defined.

Predictor Variable	Definition
Age	Age at time of injury.
Male	Being male or not being male.
HS	Had a high school diploma or equivalent or more than high school education at time of injury.
Marital Status	Being married or not married at time of injury.
Employed	Being employed full-time or not being employed full-time at time of injury.
Insurance	Having PPO or HMO insurance, insurance other than PPO or HMO, or no insurance.
CHI	Incurred a closed-head injury or other type of brain injury.
LOSIPR	Length of stay for inpatient rehabilitation hospitalization; up to three separate admissions summed in days.
DOCS-Average	The sum of each participant DOCS measures divided by the total number of DOCS evaluations [$\sum(\text{DOCS-1} + \dots + \text{DOCS-6})/\text{No. DOCS evaluations}$]; average DOCS measure.
DOCS-1	Initial DOCS neurobehavioral measure; DOCS measure from first DOCS evaluation; baseline DOCS.
DOCS-1 Days	Number of days after injury that DOCS-1 was obtained.
DOCS-Slope	DOCS Neurobehavioral Recovery Slope (β_1) as derived from mixed random effects regression analyses of 95 participants (68 CHIs; 27 other types of brain injuries).
DOCS-Intercept	DOCS initial severity level (β_0) as derived from mixed random effects regression analyses of 95 participants (68 CHIs; 27 other types of brain injuries).

DOCS = Disorders of Consciousness Scale, PPO = preferred provider organization, HMO = health maintenance organization, CHI = closed-head injury, HS = high school

individual participant's initial severity ($p = 0.001$) and rate of improvement between individual participants ($p = 0.04$). No significant covariance was found between these two terms ($p = 0.07$).

DOCS-1 Time After Injury: Subsamples

Follow-up data were collected for 72 of the 95 participants. Bivariate and multivariate analyses include all DOCS-1 assessments regardless of when the first DOCS assessment was completed (8 to 424 days of injury). Bivariate and multivariate analyses were then repeated on a subsample of 55 participants (55/72) who received the DOCS-1 assessment within 94 days of injury.

Bivariate Results and Multivariate Model Development

We used chi-square tests (or Fisher's Exact test), t -tests, and Pearson correlation coefficients for bivariate analyses evaluating the association between predictor variables and the recovery of consciousness at 1 year. Results guided development of logistic regression models used to examine multiple predictor variables for recovering consciousness. Correlations between all variables were examined pairwise. We avoided instability in

the logistic regression model estimator by not including variables in the model together if they had correlations greater than 0.70 (i.e., DOCS-Slope/DOCS-Average, $r = -0.83$; DOCS-Intercept/DOCS-Average, $r = 0.91$; DOCS-Average/DOCS-1, $r = 0.84$; DOCS-Slope/DOCS-Intercept, $r = -0.94$; DOCS-1/DOCS-Intercept, $r = 0.82$; and DOCS-1/DOCS-Slope, $r = -0.87$).

Bivariate results of the 72 participants are as expected and indicate that persons who recovered consciousness within 1 year had a significantly higher percentage of CHIs, had significantly better DOCS-Intercept and DOCS-Average, were seen for their first DOCS assessment significantly earlier after injury, and had significantly longer length of stay for IP rehabilitation (LOSIPR) (**Table 5**). DOCS-1 measures were better (higher), but only at the trend level (i.e., p -value between 0.051 and 0.10) for those recovering consciousness 1 year after injury.

Bivariate analyses of the subsample of 55 participants who had a DOCS-1 administered within 94 days of injury indicate that persons who recovered consciousness had significantly higher (better) DOCS-Average, DOCS-Slope, and DOCS-Intercept. The DOCS-Slope, an indicator of

recovery rate, is significantly different between those who recover and those who do not recover consciousness within 1 year of injury when obtained from DOCS measures obtained before 94 days of injury. DOCS-1, LOSIPR, and percent with CHI were higher at the trend level.

Predictive Values Positive and Negative and Multivariate Model Development

A receiver-operating characteristic (ROC) curve was constructed for the subsample of 55 persons first evaluated with the DOCS within 94 days of injury (**Appendix, Figure**, available in online version only). We used 10, 25, 50, 75, and 90 percent quintiles of the DOCS-1 as cut points to compare the predicted recovery with the actual recovery. The corresponding true positive and false positive rates are summarized in **Table 6**. The median DOCS-1 cut point (48.08) is the most balanced with initial DOCS

accurately predicting the recovery of consciousness 71 percent of the time and the lack of recovery 68 percent of the time. The area under the ROC curve is 0.73, indicating that the DOCS-1 can discriminate between persons who did and did not recover consciousness within 1 year 73 percent of the time.

Multivariate Results: Predicting Recovery of Consciousness up to 1 Year After Injury

We used the SAS[®] (Statistical Analysis System) procedure LOGISTIC to conduct the modeling and initially fit the logistic regression model, including all predictor variables. A stepwise procedure was employed with the use of backward elimination. The model routine ceased removing variables when no variable had a significance level greater than 0.05. To allow for the possibility of a

Table 5.

Bivariate analyses according to entire sample and subsamples (mean \pm SD).

Predictor Variable	Total Sample (DOCS-1 = 8–424 days after injury; $n = 72$)			Subsample (DOCS-1 = 8–94 days after injury; $n = 55$)		
	Recovered Consciousness Within 365 Days ($n = 46$)	Did NOT Recover Consciousness Within 365 Days ($n = 26$)	p -Values	Recovered Consciousness Within 365 Days ($n = 38$)	Did NOT Recover Consciousness Within 365 Days ($n = 17$)	p -Values
Age	34.5 \pm 15.1	34.6 \pm 12.5	0.97	33.4 \pm 15.9	36.9 \pm 13.3	0.43
Male	89.1%	76.9%	0.19	89.5	76.5	0.24
HS	31.8%	40.9%	0.59	30.6	33.3	0.99
Marital Status	39.1%	48.0%	0.62	36.8	52.9	0.38
Employed	54.6%	52.0%	0.99	50.0	58.8	0.57
Insurance	59.5%	60.0%	0.99	57.1	76.9	0.32
CHI	80.4%	53.9%	0.03*	84.2	58.8	0.08
LOSIPR	65.7 \pm 36.0	39.7 \pm 38.4	0.01*	67.2 \pm 35.5	45.5 \pm 45.6	0.08
DOCS-Average	0.97 \pm 1.1	0.04 \pm 1.3	0.002*	0.96 \pm 1.1	-0.24 \pm 1.1	0.0004*
DOCS-1	0.18 \pm 1.3	-0.53 \pm 1.2	0.06	0.18 \pm 1.3	-0.53 \pm 1.3	0.06
DOCS-1 Days	66 \pm 56	106 \pm 92	0.05*	47 \pm 22	54 \pm 19	0.24
DOCS-Slope	0.07 \pm 0.12	0.11 \pm 0.14	0.19	0.06 \pm 0.13	0.14 \pm 0.13	0.04*
DOCS-Intercept	-0.27 \pm 1.0	-0.91 \pm 1.1	0.002*	-0.19 \pm 0.9	-1.06 \pm 1.0	0.002*

*Significantly different with two-tailed $\alpha = 0.05$.

CHI = closed-head injury, DOCS = Disorders of Consciousness Scale, HS = high school, LOSIPR = length of stay for inpatient rehabilitation, SD = standard deviation

Table 6.

Predictive values positive and negative.

DOCS-1 Cut Point (DOCunits)	True Positive (%)	True Negative (%)	False Positive (%)	False Negative (%)	Correctly Classified (%)
30.32	18	95	5	82	71
42.2	41	84	16	59	71
48.08	71	68	32	29	69
53.84	82	37	63	18	51
63.92	88	13	87	12	36

DOCS = Disorders of Consciousness Scale

different pattern of recovery for CHI versus other BI, we adjusted logistic regression models for that factor.

We conducted two logistic regression models for recovering consciousness within 1 year using the DOCS-1 in continuous form and dichotomizing at median value 48.08 DOCunits. We completed the latter to assure linearity in the DOCunit. As a continuous predictor, the DOCS-1 reached trend level ($p = 0.07$), with an estimated odds ratio of 1.272 for each 50 DOCunits (95% CI; 0.97, 1.66). This indicates that for two persons assessed with the DOCS for the first time within 94 days of injury and with comparable characteristics, the person with a DOCS-1 measure 50 DOCunits higher is 1.3 times more likely to recover consciousness by 1 year.

As a dichotomized predictor, the DOCS-1 was highly significant ($p = 0.01$), with an estimated odds ratio of 0.2 (95% CI; 0.06, 0.67). This ratio indicates that for two participants assessed with the DOCS for the first time within 94 days of injury and with comparable characteristics, the participant with DOCS-1 greater than 48.08 is five times ($1/0.20$) (c-index = 0.70) more likely to regain consciousness in 1 year than the participant with a DOCS-1 less than 48.08 (Figure 5). The participant with a DOCS-1 less than 48.08 has about a 60 percent chance of recovering consciousness 210 days after injury. A participant with a DOCS-1 ≥ 48.08 has a 60 percent chance of recovering consciousness 112 days after injury and about a 90 percent chance of recovering consciousness 182 days after injury.

When the dichotomized DOCS-1 is modeled with the covariates, none is a significant predictor in addition to the dichotomized DOCS-1 measure. However, controlling for CHI, we found that the DOCS-1 continues signifi-

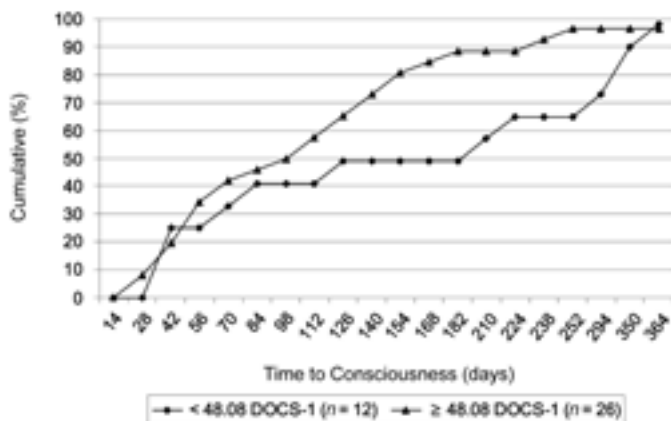


Figure 5. Probability of recovering consciousness according to days after injury.

cantly to predict recovery of consciousness up to 1 year after injury ($p = 0.02$). We then computed the odds ratio as 0.23 (95% CI; 0.06, 0.85), controlling for CHI, indicating that a participant with a DOCS-1 ≥ 48.08 is about four times more likely to recover consciousness in one year than a participant with a DOCS-1 less than 48.08. This model was able to distinguish between patients who regained consciousness and those who did not 73 percent of the time (c-index = 0.73).

A fourth model was fit with the use of the DOCS-Average and variables that would influence recovery. Three variables significantly predicted the recovery of consciousness 1 year after injury: DOCS-Average ($p = 0.02$), CHI ($p = 0.03$), and LOSIPR dichotomized at 28 days ($p = 0.001$). The estimated odds ratio for the DOCS-Average is 1.4 per 50 DOCunit change (95% CI; 1.1, 2.0), for CHI is 4.3 (95% CI; 1.1, 16.7), and LOSIPR > 28 days is 10.8 (95% CI; 2.6, 45.9). The model could correctly classify whether or not a patient regained consciousness 86 percent of the time (c-index = 0.86).

DISCUSSION

Findings indicate that the DOCS, when comprising 23 test stimuli, produces a measure that is a reliable and valid indicator of subtle changes in neurobehavioral functioning in unconscious persons over time. Findings also indicate that the DOCS can accurately predict the recovery of consciousness up to 1 year after injury. The abbreviated DOCS comprises a common set of 23 test stimuli that allied health professionals can score and administer within 15 to 30 minutes as early as 8 days after injury to persons who are unconscious following a CHI, an anoxic event, an aneurysm, an open-head injury, and/or a hemorrhagic event. This is the first time in published literature that a neurobehavioral measure has been reported to predict the recovery of consciousness up to 1 year after injury and used with multiple etiological groups.

The variables found to predict the recovery of consciousness 1 year after injury include the initial DOCS measure (DOCS-1) obtained within 94 days of injury, DOCS-Average, LOSIPR, and CHI. Two of the three logistic regression models were significant ($p < 0.05$), and the third was significant at the trend level. The model with the lowest predictive certainty (70%) used a dichotomized DOCS-1 to distinguish between high and low performers and controlled for etiology. The model with

